2019 State Damage Prevention Program Grants Final Report
CFDA Number: 20.720

Award Number: 693JK31940021PSDP
Project Title: State Damage Prevention (SDP) Program Grants - 2019
Date Submitted: *November 17, 2020*
Submitted by *Scott Crawford*

Specific Objective(s) of the Agreement

Under this grant agreement, the recipient will:

Identify a statistically valid sampling of tickets to audit, conduct a statistic audit of its current auditing process, and create an algorithm that can identify high-risk tickets that can then be audited, with the algorithm being capable of adjusting weights of variables based on data gathered from audits. (Elements 1-9)

Work Scope:

Under the terms of this grant agreement, the Recipient will address the following applicable elements listed in the approved application, pursuant to 49 U.S.C. §60134 (a), (b).

- Element 1 (Effective Communications): Participation by operators, excavators, and other stakeholders in the development and implementation of methods for establishing and maintaining effective communications between stakeholders from receipt of an excavation notification until successful completion of the excavation, as appropriate. (Applicable)
- Element 2 (Comprehensive Stakeholder Support): A process for fostering and ensuring the support and partnership of stakeholders, including excavators, operators, locators, designers, and local government in all phases of the program. (Applicable)
- Element 3 (Operator Internal Performance Measurement): A process for reviewing the adequacy of a pipeline operator's internal performance measures regarding persons performing locating services and quality assurance programs. (Applicable)
- Element 4 (Effective Employee Training): Participation by operators, excavators, and other stakeholders in the development and implementation of effective employee training programs to ensure that operators, the one call center, the enforcing agency, and the excavators have partnered to design and implement training for the employees of operators, excavators, and locators. (Applicable)
- Element 5 (Public Education): A process for fostering and ensuring active participation by all stakeholders in public education for damage prevention activities. (Applicable)
- Element 6 (Dispute Resolution): A process for resolving disputes that defines the State authority's role as a partner and facilitator to resolve issues. (Not Applicable)
- Element 7 (Enforcement): Enforcement of State damage prevention laws and regulations for all aspects of the damage prevention process, including public education, and the use of civil penalties for violations assessable by the appropriate State authority. (Not Applicable) •
- Element 8 (Technology): A process for fostering and promoting the use, by all appropriate stakeholders, of improving technologies that may enhance communications,

underground pipeline locating capability, and gathering and analyzing information about the accuracy and effectiveness of locating programs. (Applicable)

- Element 9 (Damage Prevention Program Review): A process for review and analysis of the effectiveness of each program element, including a means for implementing improvements identified by such program reviews. (Applicable)

Accomplishments for the grant period (Item 1 under Agreement Article IX, Section 9.02 Final Report: "A comparison of actual accomplishments to the objectives established for the period.")

Virginia Utility Protection Service, Inc. completed all objectives identified for the grant. The following list denotes accomplishments for this grant:

1. Vendor selected
    a. Statistical Applications & Innovations Group (SAIG), Virginia Tech
2. Contracts signed
    a. Executed Professional Services Agreement with Vendor
    b. Agreement for Disclosure and Transfer of Confidential Information and Personally Identifiable Information
3. Completed the Statement and Scope of Work with SAIG
4. Created and delivered to vendor the Data Dictionary
5. Delivered to vendor three years of web ticket entry audit data fields to be used in development of AI
6. Vendor created and delivered the VA811 Safety Risk Rating Auditor Agreement
7. Worked with vendor on developing the VA811 Safety Risk Rating Application
8. Tested the application
9. Created a Metrics report from use of the application

1. Purpose

Phase I of the project consisted of assessing, using statistical modeling, the overall effectiveness of VA811's current auditing procedures in order to determine the current level of auditing effectiveness based on three criteria: 1) auditing repeatability (degree to which same auditors examining same tickets get the same results); 2) auditing reproducibility (degree to which different auditors examining the same tickets get the same results); 3) auditing accuracy (degree to which auditors achieve the same results auditing tickets as did the experts). A fourth component relates to analyzing whether the three areas related to repeatability, reproducibility, and accuracy statistically improve as the current 25 audit codes are reduced to 13 merged codes, 4 ordinal codes, and a binary assessment. Phase I is also involved SAIG identifying a statistically valid sampling for random auditing purposes.

Phase II took the Data Dictionary and the ticket audit data, along with information gleaned from the Phase I deliverables, to create a learning algorithm to ensure 100% auditing of web tickets.

2. Methodology

A total of 25 Damage Prevention Specialists (DPS) involved in the auditing process examined and, using one of 25 audit codes and "verified," indicating the ticket presented no evidence of error, scored 50 tickets. A team of 4 "experts" created the scoring key, with the key identifying the accurate code for each ticket or determining the ticket was verified. Three of the 25 DPS were randomly chosen to audit the same tickets again roughly two and a half weeks after the initial auditing of the tickets. Through this process, SAIG was able to statistically analyze the results for accuracy, repeatability, and reproducibility.

3. Value

The completion of Phase I provided VA811 with valuable insights into its auditing process. It is hypothesized that reducing the auditing process to a binary classification will increase overall auditing accuracy. The auditing process will involve the wider pool of auditors simply classifying tickets as verified, meaning no evidence of error, or as containing a Safety Level, meaning evidence of error exists. Tickets identified as possibly containing an error is then be turned over to a smaller and dedicated QA/QC team to assign an agreed upon audit code or determine the ticket is verified. Upon completion of the identification of a statistically valid random number of tickets to audit based on ticket volume and error rate, VA811 adjusted its current auditing practices to audit the identified number of tickets. At the conclusion of Phase II, VA811 has begun to work with its software development software company to put in place the learning algorithm so that 100% of web-originated normal tickets will be screened using the binary classification system. Any tickets the AI engine (learning algorithm) identifies as possibly containing an error will be audited by the dedicated QA/QC team. DPS auditors will continue to also audit a statistically valid random sampling of DPS originated tickets using this binary classification system, sending tickets with potential Safety Level concerns to the QA/QC team.

See Appendix A – VA811 Safety Risk Rating Auditor Agreement
See Appendix B – Gradient Boosting Machine Model for Predicting Safety Violations

Quantifiable Metrics/Measures of Effectiveness (Item 2 under Article IX, Section 9.02 Final Report: "Where the output of the project can be quantified, a computation of the cost per unit of output.")

See Appendix C – Quantifiable Metrics Report

Issues, Problems or Challenges (Item 3 under Article IX, Section 9.02 Final Report: "The reasons for slippage if established objectives were not met. ")

All objectives were met. VA811 did request and receive approval for late submission of the final report due to the impact of the COVID19 virus on vendor staffing.

Final Financial Status Report

Deliverable	Price	Object Class Category
Research & Development	$10,69.76	Contractual
Research & Development	$16,638.20	Contractual
Research & Development	$ 5,187.18	Contractual
Research & Development	$21,559.80	Contractual
Research & Development	$15,448.62	Contractual
Research & Development	$11,883.70	Contractual
Research & Development	$17,970.61	Contractual

See Appendix D –Invoices/Check for services/payments and SF 425.

Requests of the AOR and/or PHMSA

No actions requested at this time.

Appendix A
VA811 Safety Risk Rating Auditor Agreement

VIRGINIA TECH

SAIG
Data • Science • Decisions
Statistical Applications & Innovations Group

VA811 Safety Risk Rating Auditor Agreement :

Accuracy, Reproducibility & Repeatability

Jennifer H Van Mullekom, PhD
Eric Bae, MS
VIRGINIA TECH | STATISTICAL APPLICATIONS & INNOVATIONS GROUP | MARCH 25, 2020

VA811 Safety Risk Rating Auditor Agreement: Accuracy, Reproducibility & Repeatability

Executive Summary

This report assesses the initial screening measurement system of web excavation ticket entry at VA811 for safety level risk. Accuracy, repeatability, and reproducibility of the auditors on an 25 audit code measurement scale is characterized via attribute agreement analysis, a form of measurement systems analysis for discrete data. Three additional scales which collapse the 25 code scale into smaller numbers of categories were also assessed. These analyses were performed for two purposes: 1) providing information on operations improvement for VA811 and 2) determining which scale should be used as the dependent variable for a predictive model to inform safety level risk audits.

Statistical analyses indicate the measurement system with 25 three digit safety risk codes has poor accuracy (auditors have poor agreement with experts), poor repeatability (auditors have poor agreement with themselves), and poor reproducibility (auditors have poor agreement with each another). When the scales are collapsed reducing the number of codes, the quality of the measurement system improves. However, additional steps should be taken to ensure a fully validated measurement system. This report as well as a wealth of human factors rating research supports reducing the number of codes in the initial screening audit for safety level risk. In addition, it supports the use of a two level safety risk scale (safety violation/no safety violation) for predictive modeling.

Introduction

As more customers employ VA811's web ticket systems to request utility identification prior to excavation, detection of potential errors in tickets with high accuracy is essential. The web ticketing system contains safety level risks in a higher proportion than direct calls to VA811 representatives. VA811 has contracted with VT SAIG to ultimately provide a predictive model that will aid in the identification of high risk tickets. The predictive model will be implemented as part of a new risk-based audit plan. Prior to modeling, it is important to understand the accuracy and precision of the safety level violation system that will be used in the modeling. This report details the measurement system analyses (MSA) for the audit safety level violation system on various scales derived from the original 25 category scale.

Current methods of detecting errors involve multiple auditors assigning safety level codes. Auditors assigned the status as "accurate" if no issue is found. After this initial assessment, the ticket is forwarded to an expert for assignment of a final audit code and resolution of the safety risk issue. For this MSA, 15 different auditors and a consensus panel of experts audited 50 tickets. All 15 auditors assigned one of the 25 safety level codes specified in current procedure to each of the 50 tickets on the 10th of December 2019, whereas three of the 15 auditors performed yet another audit on the 31st of December 2019.

This report will answer four questions regarding the measurement system:

1. Is the measurement system repeatable? Do the ***same*** auditors reviewing the ***same*** tickets, get the same results on multiple trials?
2. Is this measurement system reproducible? Do ***different*** auditors reviewing the ***same*** tickets get the same results?
3. Is this measurement system accurate? Do ***auditors*** get the same results as reached in the consensus session by ***experts***?
4. Do the quantities which evaluate repeatability, reproducibility, and accuracy improve as we reduce the 25 safety level codes to 13 merged codes, to 4 ordinal codes, and finally to a binary assessment (accurate versus violation)?

Details of the methodology are seen in the Methods section. Statistical tools used are explained in detail in the Analysis section. Output of the analysis is summarized in the Results section. The Appendix contains detailed output from each analysis with an explanation of interpretation in each section of the Appendix.

METHODS

The MSA will be performed using attribute agreement analysis. In this attribute agreement analysis, we start by asking ourselves the following questions:

1. Is the outcome consistent over different trials for each auditor? (Repeatability)
2. Is the outcome consistent across all auditors? (Reproducibility)
3. Is the outcome consistent for both auditors and experts? (Accuracy/Bias)
4. Is there a functional difference between leaving all categories as possible choices versus merging similar categories together?

The first question represents repeatability. Poor repeatability indicates inconsistency in individual auditors. The second question represents reproducibility. Poor reproducibility suggests that there is high variability among safety level codes assigned to the same ticket by multiple auditors. Accuracy is assessed by comparing auditors rating to the panel of experts' consensus rating. The first three questions are part of what is considered "Attribute Agreement Analysis," and involves statistical tools such as Fleiss' Kappa and Kendall's W and Tau. These tools will be further explained in the *Analysis* section. Finally, the fourth and final question asks if reducing total category options leads to improved measurement systems.

To answer the fourth question, four different scenarios of category options will be analyzed using Attribute Agreement Analysis menu option in Minitab™. The results of this analysis will be compared to each other and to benchmarks common to MSA. The scenarios of category options considered – all codes, reduced codes, risk level, and violation status—are shown in Table 1.

Table 1. **Original Scale and Modified Scales**

Original Scale	Modified Scales		
Full Codes	Similar Merged	Risk Level	Violation Status
Accurate	*Accurate*	*1*	*No*
110*	110	1	No
140	140	1	No
150*	150	1	No
151		1	No
152		1	No
161*	161	1	No
162*		1	No
170	*170*	*1*	*No*
171		1	No
172*		1	No
174		1	No
181*	181	2	Yes
190	*190*	*2*	*Yes*
191		2	Yes
240	*240*	*4*	*Yes*
241		4	Yes
260	260	4	Yes
261*	261	4	Yes
262*		4	Yes
270	*270*	*4*	*Yes*
271		3	Yes
273		3	Yes
290	*290*	4	Yes
291		*4*	*Yes*

* Codes that were not used by any of the auditors were starred

* Codes that the experts used are underlined and in italics

Table 1 illustrates how the categories were merged from one scenario to the next. There were a total of 25 different codes including "Accurate;" however, only 18 were used by any of the 50 auditors.

Repeatability and reproducibility analyses were performed directly on the data set with the original codes. After that, the analyses were performed on merged codes (second column of *Table 1*), risk levels (third column), and violation status (fourth column). Note that on the risk level scenario, the option "accurate" has been merged into codes colored green, representing "minimal risk" category. On the violation status, all options formally belonging to 110 to 174 were collapse into the "No Violation Status" or, alternatively, "Accurate" as the potential risk associated with tickets assigned to those codes are very low. The remaining codes are considered a violation.

There are two types of agreements: absolute and relative agreements. Absolute agreement requires an exact match and is most commonly used in measurement systems analysis with nominal or named categories that have no ordering. Absolute agreement percentage is the total number of tickets with agreement divided by the total number of tickets. Absolute agreement is calculated for all four

scenarios. Relative agreement, however, does put emphasis on the scale, and is used for cases with ordinal variables. As the name implies, ordinal variables have an order to them but there is not a defined numerical interpretation for the distance between categories. Low, medium, high or even 1-5 on a survey rating scale are examples of ordinal scales. Relative agreement is calculated for the risk level scenario scaled 1-4 that has an order of severity in column 3 of Table 1.

To further characterize absolute and relative agreement, consider two different auditors rating the same ticket with 151 and 152. This will still count as the same "disagreement" as rating it with 151 and 290, despite the fact that the latter appears to be a more serious disagreement. Fleiss's Kappa and the agreement percentage both measure absolute agreement. For the risk level scenario, additional statistics called Kendall's W and Kendall's Tau were calculated. Output of these statistics can be seen on *Table 2* on the Results section. The Kendall statistics measure relative agreement for ordinal data.

For the repeatability section, in addition to those on the reproducibility, agreement percentage and the kappa values were calculated for each auditor's two assignments and each auditor against the expert. *Table 3* shows the output of this.

Analysis

In this analysis, Fleiss' Kappa, Kendall's W, and Kendall's Tau were utilized to develop attribute agreement measurement systems. Fleiss' Kappa measures absolute agreement and the Kendall's W and Tau measure relative agreement for ordinal data. Minitab™ was used to calculate these statistics, and the Minitab™ output can be seen in the Appendix section.

Fleiss' Kappa

Fleiss' Kappa measures the degree of agreement over and above the amount of agreement by chance. The Kappa can take the value between -1 and 1, where 1 represents complete agreement, and -1 represents complete disagreement, and 0 represents agreement level that is equal to the level that would have been obtained completely by chance. As a rule of thumb for measurement system analysis, the Kappa value of above 0.9 qualifies as acceptable.

Kendall's W

Both Kendall's W and Kendall's Tau are applied when the outcome is ordinal. Of the four scenarios – all codes, reduced codes, risk levels, and violation status – only the risk levels involve ordinal measurements of between 1 to 4. Therefore, both W and Tau are applied only when analyzing risk levels.

Kendall's W measures the degree of association of ordinal assessments made by multiple auditors when assessing the same samples. The W can take any value between 0 and 1, where 0 represents no concordance and 1 represents perfect concordance.

KENDALL'S TAU

Kendall's Tau, also known as Kendall's correlation coefficient, is a correlation coefficient specifically for ordinal variables and, therefore, follow values between -1 and 1, where -1 represents complete opposite and 1 represents complete match.

For both Kendall's W and Tau, the same rule of thumb of above 0.9 as an acceptable outcome applies.

RESULTS

This section contains a brief summary of the results obtained from the assessment agreement analysis as laid out on the Introduction section. Repeatability and reproducibility outputs will be mentioned separately.

A heat map will be used to compare an ideal system with the observed data from this study. *Figure 1* illustrates the heat map of the assessment of the auditors against the expert. In an ideal situation, where all auditors agreed with the expert 100% of the time, the left heat map would be produced. The ideal map shows all points occupying cells in the along the diagonal from the bottom left to the top right. The right side of the figure is the heat map that was obtained from the study. While some patterns of diagonals are identifiable, it is clear that there is a distinct visible difference from the "ideal" heat map.

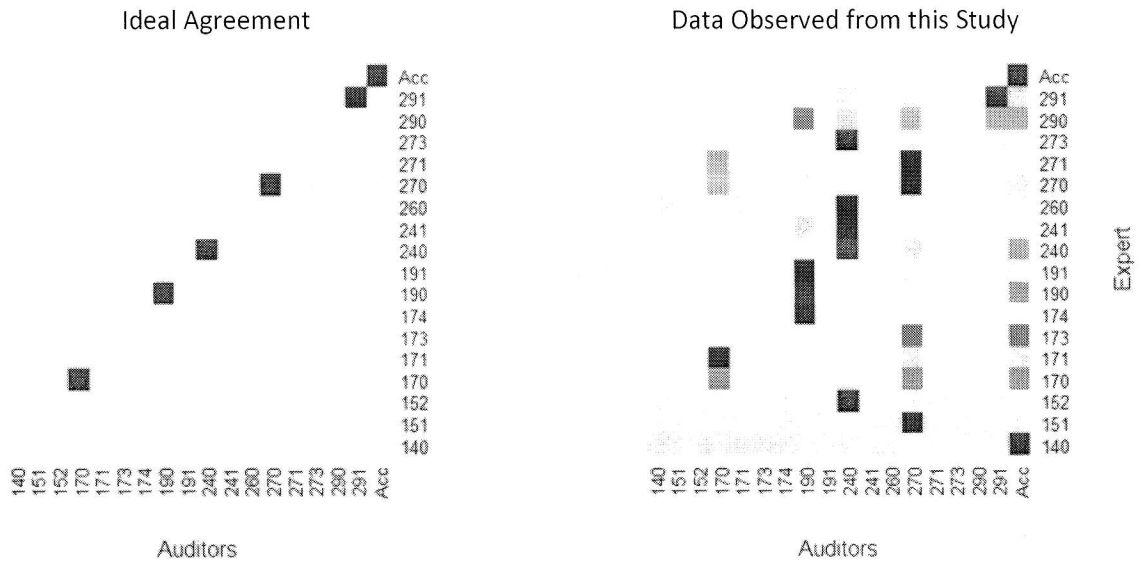


Figure 1

Left: An ideal scenario where all auditors' predictions matches with the expert's
Right: Current scenario

Table 2 is a summary of the reproducibility and accuracy. The measures of agreement increase moving from the left column to the right column of the table. The more the scale (codes) are collapsed, the higher the average percentage agreement and the higher the average Kappa value for both reproducibility and accuracy. Likewise, the percentage of unanimous agreement for both among the 15 auditors and with the expert increased as more codes were merged.

However, both the percent of agreement and the overall Kappa appears to remain far below the acceptable values of 90% and 0.90 respectively, as do the Kendall statistics. Even the last scenario considered, violation status as a binary outcome, representing the least complicated scale does not achieve this benchmark. This prompts consideration for a measurement system improvement project.

Table 2 Reproducibility and Accuracy summary table of the relevant output of the four scenarios

		Full Codes	Similar Merged	Risk Level	Violation Status
Reproducibility Between-auditors	% Agreement Range (% Average)	36 – 64 (52.80)	42 – 70 (58.40)	58 – 76 (67.87)	62 – 82 (73.60)
	Kappa Range (Overall Kappa)	<0 – 0.47 (0.29)	<0 – 0.47 (0.36)	0.17 – 0.43 (0.40)	0.43 (0.43)
	Unanimous % Agreement	6	8	24	28
	Kendall's Tau	-	-	0.52	-
Accuracy Auditor vs. Expert	Kappa Range vs. Expert (Overall Kappa)	0.20 – 0.58 (0.40)	0.22 – 0.57 (0.45)	0.47 – 0.48 (0.46)	0.47 (0.47)
	Unanimous % Agreement vs. Expert	6	8	24	28
	Kendall's W (Overall)	-	-	0.24 – 0.61 (0.47)	-

Table 3 is the summary table for repeatability measurements. There were three auditors – auditors 9, 10, and 11 – who evaluated the sample of 50 tickets twice – once at the 10th of December, 2019 and another at the 31st of December, 2019. The same metrics described in Table 2 are reported in Table 3. Once again, with fewer categories, we observe higher agreement range among the auditors themselves, each auditor against the expert, and all three auditors and the expert.

Note that the percentage of agreement differs when it is calculated among auditors and when the auditors were compared to the experts. This is because all auditors agreed unanimously on certain tickets but did not agree with the experts.

Table 3. Repeatability summary table of the relevant output of the four scenarios

		Full Codes	Similar Merged	Risk Level	Violation Status
Repeatability Auditor vs. Self	% Agreement Range (% Average)	54 – 72	60 – 86	68 – 82	78 – 90
	Kappa Range (Overall Kappa)	<0 – 1* (0.34 – 0.64)	<0 – 1* (0.44 – 0.81)	<0 – 0.82 (0.33 – 0.78)	0.48 – 0.80
	Kendall's Tau	-	-	0.76-0.89	-

In every scenario, for both reproducibility and repeatability, the goal of over 90 % average prediction accuracy rate and 0.90 Fleiss' Kappa, Kendall's W, and Kendall's Tau was not met.

DISCUSSION

Accuracy

The ability of the auditors to align their classifications to the expert consensus classification falls below the desired thresholds of 90% for absolute agreement and 0.9 for Kappa or Kendall's statistics.

Reproducibility

The reproducibility analysis shows that as more codes were merged, agreement increased among the auditors. Fleiss' Kappa statistics also improved. However, in all of the four scenarios, the improvement in accuracy fell short of our standard of 90% for absolute agreement and 0.90 in Kappa and Kendall's statistics. This is an indication that reproducibility should be improved significantly.

Repeatability

The repeatability analysis also shows that as more codes were merged, agreement between each auditor's assessments on multiple trials improved. However, similarly to the reproducibility result, improvements as the scale was collapsed usually fell well short of the aforementioned standard. There were a few exceptions with one auditor into the ordinal risk level 1-4 measurement and the binary violation status measurement.

CONCLUSIONS AND RECOMMENDATIONS

The analyses in this report were performed for two purposes: 1) providing information on operations improvement for VA811 and 2) determining which scale should be used as the dependent variable for a predictive model to inform safety level risk audits.

With respect to operations improvement, this study shows that the number and complexity of audit code descriptions impair accuracy, repeatability, and reproducibility. While collapsing categories suggests improvement in the system, this was done via computer during data analysis. The analysis indicates a trend but does not necessarily characterize the full potential of such a shift to a scale with fewer categories. It is hypothesized that a measurement system with as possible would provide a marked improvement over the derived 4 level risk scale and 2 category binary scale analyzed in this study. Such a transition must be accompanied by thorough operational definitions and appropriate training with a follow up MSA. This transition is supported by human factors research summarized in the following quote: "In general, inspection performance is degraded as the number and types of defects increases, primarily as a result of limitations of human memory. " (Dalton & Drury, 2004). For a thorough consideration of all factors in the design of a human visual inspection system, see https://prod-ng.sandia.gov/techlib-noauth/access-control.cgi/2012/128590.pdf

An initial ticket audit is used to send the excavation ticket through final audit by supervisors and experts. The final audit codes recorded for tickets reflect expert opinion. All 25 detailed codes are eligible to be recorded for a final classification by the expert auditor. This final classification will be modeled in the second phase of the project. Based on this study, we propose converting the 25 code scale to a binary classification of (no violation, violation) for modeling phase of the project. This decision is based on both the quality of the measurement system and types of models planned for subsequent phases of the project.

References

Sandia Literature Review of Visual Inspection: https://prod-ng.sandia.gov/techlib-noauth/access-control.cgi/2012/128590.pdf

Dalton, J., & Drury, C.G. (2004). Inspectors' performance and understanding in sheet steel inspection. Occupational Ergonomics, 4, 51-65.

APPENDIX

This section contains an overview of a general interpretation of each item of the analysis on *Tables 2* and *3* (first and second columns). Most output has been generated using Minitab™. For each part, an example output is provided. Unless otherwise stated, all examples provided are extracted from output produced using the full 3-digit codes. The Minitab™ File with raw data will be provided as part of the documentation package.

Reproducibility

For the reproducibility, there were five key parts – assessment agreement percentage, Kappa statistic within auditors, unanimous agreement amongst auditors, Kappa statistic between auditors and the experts, and the unanimous agreement between the auditors and the experts.

Between-auditors, Assessment Agreement

The assessment agreement is obtained by comparing each auditors' assessment to that of the experts' and tallying up the percentage of assessments that matched. Figure 2 illustrates an example of the assessment agreement of each of the 15 auditors, along with a 95 % confidence interval. This figure is generated from when all three-digit codes were used.

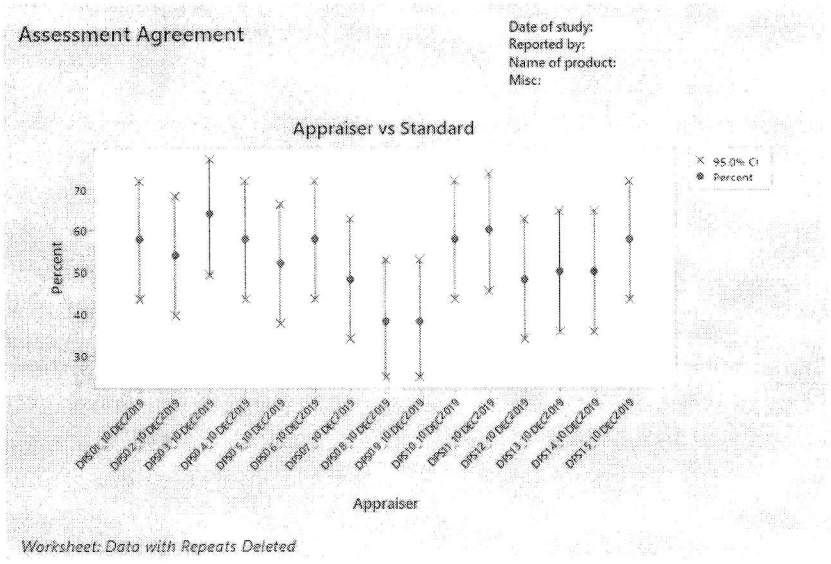


Figure 2 Example of an assessment agreement plot

Between-auditors, Kappa Range

Fleiss' Kappa statistic has been described in the Analysis section. This Kappa statistic calculates the auditors' degree of agreement on each of the responses. A sample output is in Figure 3, which illustrates the Kappa statistics of all 3-digit codes that were used by any of the auditors.

Fleiss' Kappa Statistics

Response	Kappa	SE Kappa	Z	P(vs > 0)
140	-0.0054	0.0138	-0.3885	0.6512
151	-0.0013	0.0138	-0.0967	0.5385
152	0.2809	0.0138	20.3546	0.0000
170	0.1409	0.0138	10.2063	0.0000
171	0.1359	0.0138	9.8501	0.0000
173	-0.0027	0.0138	-0.1937	0.5768
174	0.3520	0.0138	25.5018	0.0000
190	0.2046	0.0138	14.8226	0.0000
191	0.1798	0.0138	13.0290	0.0000
240	0.3155	0.0138	22.8584	0.0000
241	0.1659	0.0138	12.0183	0.0000
260	0.1394	0.0138	10.1016	0.0000
270	0.2923	0.0138	21.1819	0.0000
271	0.1789	0.0138	12.9634	0.0000
273	-0.0013	0.0138	-0.0967	0.5385
290	0.0623	0.0138	4.5114	0.0000
291	0.4287	0.0138	31.0622	0.0000
Acc	0.4702	0.0138	34.0672	0.0000
Overall	0.2949	0.0060	49.2635	0.0000

Figure 3 Fleiss' Kappa statistic by each response

This example shows that none of the responses reached the Kappa value of 0.5, let alone 0.9, which indicate poor level of agreement among auditors. There are, likewise, a few with Fleiss' Kappa smaller than zero, though none of them appeared to have P-value (far right column) small enough (under 0.05) to be considered significantly different from zero.

Between-auditors, unanimous agreement

This value represents the percentage of tickets on which all 15 auditors agreed to a particular code value. The figure below indicates that, when none of the 3-digit codes were merged into common categories, of the 50 tickets analyzed, all auditors agreed on 3, which corresponds to 6 percent of all tickets.

Assessment Agreement

# Inspected	# Matched	Percent	95% CI
50	3	6	(1.25, 16.55)

 # Matched: All appraisers' assessments agree with each other.

Figure 4 Total number and percentage of tickets with unanimous agreement amongst all auditors

Each auditor vs. Experts, assessment agreement

The assessment agreement is obtained by comparing each auditors' assessment to that of the experts' and tallying up the percentage of assessments that matched. Figure 2 illustrates an example of the assessment agreement of each of the 15 auditors, along with a 95 % confidence interval. This figure is generated from when all three-digit codes were used.

Assessment Agreement

Appraiser	# Inspected	# Matched	Percent	95% CI
DPS01_10DEC2019	50	29	58	(43.2, 71.8)
DPS02_10DEC2019	50	27	54	(39.3, 68.2)
DPS03_10DEC2019	50	32	64	(49.2, 77.1)
DPS04_10DEC2019	50	29	58	(43.2, 71.8)
DPS05_10DEC2019	50	26	52	(37.4, 66.3)
DPS06_10DEC2019	50	29	58	(43.2, 71.8)
DPS07_10DEC2019	50	24	48	(33.7, 62.6)
DPS08_10DEC2019	50	19	38	(24.7, 52.8)
DPS09_10DEC2019	50	19	38	(24.7, 52.8)
DPS10_10DEC2019	50	29	58	(43.2, 71.8)
DPS11_10DEC2019	50	30	60	(45.2, 73.6)
DPS12_10DEC2019	50	24	48	(33.7, 62.6)
DPS13_10DEC2019	50	25	50	(35.5, 64.5)
DPS14_10DEC2019	50	25	50	(35.5, 64.5)
DPS15_10DEC2019	50	29	58	(43.2, 71.8)

 # Matched: Appraiser's assessment across trials agrees with the known standard.

Figure 5 Total percentage of tickets for which each auditor's decision matched with that of the experts'

All auditors vs. Experts, Kappa statistic

This represents the Kappa statistics for the codes that were used by the experts. As the figure below illustrates, there were only six codes used by the experts; hence, most responses do not have Kappa statistics assigned. Also noticeable is that the Kappa values are higher for "Acc" and "291" and lower for codes in between. This is an indication that the auditors were, in general, more likely to assign codes to tickets with either no detectable risk (Acc) or with very high risk (291) but were more likely to disagree on tickets with medium-level risks.

Fleiss' Kappa Statistics

Response	Kappa	SE Kappa	Z	P(vs > 0)
140	*	*	*	*
151	*	*	*	*
152	*	*	*	*
170	0.2013	0.0365	5.5121	0.0000
171	*	*	*	*
173	*	*	*	*
174	*	*	*	*
190	0.3358	0.0365	9.1972	0.0000
191	*	*	*	*
240	0.2856	0.0365	7.8225	0.0000
241	*	*	*	*
260	*	*	*	*
270	0.3968	0.0365	10.8657	0.0000
271	*	*	*	*
273	*	*	*	*
290	*	*	*	*
291	0.5896	0.0365	16.1463	0.0000
Acc	0.5562	0.0365	15.2315	0.0000
Overall	0.3868	0.0177	21.7958	0.0000

* When all sample standards and responses of a trial(s) equal the value or none of them equals the value, kappa cannot be computed.

Figure 6 Fleiss' Kappa statistics of experts' choices against those of the 15 auditors

All auditors vs. Experts, assessment agreement

This item represents the total number of tickets on which all auditors AND the experts agreed. There were a total of three tickets (6 % of all tickets) on which all auditors and the experts agreed. This number matches with that obtained between-auditors, indicating that there were no tickets on which the auditors unanimously agreed but the experts did not.

Assessment Agreement

# Inspected	# Matched	Percent	95% CI
50	3	6	(1.25, 16.55)

Matched: All appraisers' assessments agree with the known standard.

Figure 7 Assessment agreement of all auditors and the experts

Repeatability

There were four types of assessment agreement percentage and four types of Fleiss' Kappa statistics – auditors vs. self, among auditors, each auditor vs. experts, and all auditors vs. experts. In addition, for the risk level, two Kendall's statistics were calculated as well. There were three auditors with separate assessments on the same 50 tickets on two different dates.

Auditors vs. self & each auditor vs experts, assessment agreement

The figure below shows two types of assessment agreements – within appraisers and each appraiser against the experts. The former is represented by the first plot, while the latter is represented by the second plot.

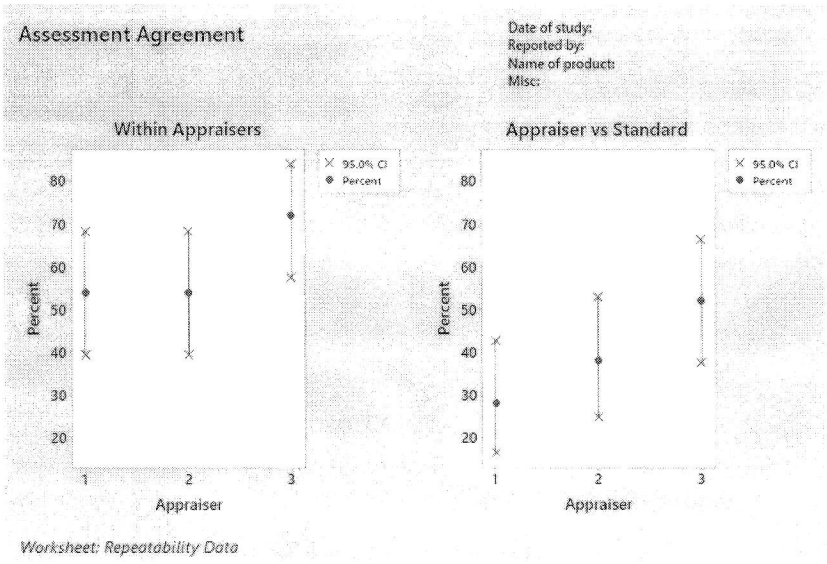


Figure 8 Assessment agreement within appraisers and each appraiser against the experts

Auditors vs. self, Fleiss' Kappa

This large table shows the Fleiss' Kappa statistics of the codes used by each of the three auditors against oneself. Ideally all responses should have Kappa statistics of 1, meaning all auditors selected the same codes for both times. In this example, there were two 1's – 152 for the auditor 1 (9) and 152 for the auditor 2 (10). However, this is mostly due to small sample size. Most others did not come near the benchmark of 0.9; this is an indication that the auditors did not have good assessment agreement level with themselves when assessed at two different times.

Fleiss' Kappa Statistics

Appraiser	Response	Kappa	SE Kappa	Z	P(vs > 0)
1	140	-0.02041	0.141421	-0.14431	0.5574
	152	1.00000	0.141421	7.07107	0.0000
	170	0.46524	0.141421	3.28975	0.0005
	171	*	*	*	*
	173	-0.01010	0.141421	-0.07142	0.5285
	174	-0.01010	0.141421	-0.07142	0.5285
	190	0.55556	0.141421	3.92837	0.0000
	191	-0.02041	0.141421	-0.14431	0.5574
	240	0.65636	0.141421	4.64115	0.0000
	241	*	*	*	*
	260	-0.01010	0.141421	-0.07142	0.5285
	270	0.34211	0.141421	2.41905	0.0078
	271	-0.01010	0.141421	-0.07142	0.5285
	290	-0.02041	0.141421	-0.14431	0.5574
	291	1.00000	0.141421	7.07107	0.0000
	Acc	0.45098	0.141421	3.18891	0.0007
	Overall	0.41846	0.066304	6.31117	0.0000

Appraiser	Response	Kappa	SE Kappa	Z	P(vs > 0)
2	140	*	*	*	*
	152	1.00000	0.141421	7.07107	0.0000
	170	0.45652	0.141421	3.22810	0.0006
	171	0.36842	0.141421	2.60513	0.0046
	173	-0.01010	0.141421	-0.07142	0.5285
	174	*	*	*	*
	190	-0.07527	0.141421	-0.53223	0.7027
	191	-0.01010	0.141421	-0.07142	0.5285
	240	-0.03093	0.141421	-0.21869	0.5866
	241	-0.01010	0.141421	-0.07142	0.5285
	260	*	*	*	*
	270	-0.07527	0.141421	-0.53223	0.7027
	271	-0.03093	0.141421	-0.21869	0.5866
	290	-0.05263	0.141421	-0.37216	0.6451
	291	0.47917	0.141421	3.38822	0.0004
	Acc	0.63870	0.141421	4.51629	0.0000
	Overall	0.33698	0.064152	5.25293	0.0000

Appraiser	Response	Kappa	SE Kappa	Z	P(vs > 0)
3	140	*	*	*	*
	152	*	*	*	*
	170	0.81061	0.141421	5.73185	0.0000
	171	-0.01010	0.141421	-0.07142	0.5285
	173	*	*	*	*
	174	-0.01010	0.141421	-0.07142	0.5285
	190	0.67511	0.141421	4.77377	0.0000
	191	-0.02041	0.141421	-0.14431	0.5574
	240	-0.02041	0.141421	-0.14431	0.5574
	241	-0.01010	0.141421	-0.07142	0.5285
	260	*	*	*	*
	270	0.76471	0.141421	5.40729	0.0000
	271	-0.01010	0.141421	-0.07142	0.5285
	290	-0.03093	0.141421	-0.21869	0.5866
	291	-0.04167	0.141421	-0.29463	0.6159
	Acc	0.87390	0.141421	6.17938	0.0000
	Overall	0.63693	0.068212	9.33748	0.0000

* When no or all responses across trials equal the value, kappa cannot be computed.

Figure 9 Fleiss' Kappa statistics of the codes by each auditor

Among auditors, assessment agreement

The figure below is a cross-auditor comparison of assessment agreement. Out of 50 tickets inspected, only five (10 % of all tickets) were unanimously agreed upon by the three auditors in BOTH dates.

Assessment Agreement

# Inspected	# Matched	Percent	95% CI
50	5	10.00	(3.33, 21.81)

 # Matched: All appraisers' assessments agree with each other.

Figure 10 Assessment agreement among auditors

Among auditors, Fleiss' Kappa

This is the Fleiss' Kappa statistics of all three auditors for both dates. Therefore, this Kappa statistic would be comparing a total of six different input for each ticket.

Fleiss' Kappa Statistics

Response	Kappa	SE Kappa	Z	P(vs > 0)
140	-0.0067	0.0365	-0.1838	0.5729
152	0.5946	0.0365	16.2836	0.0000
170	0.1509	0.0365	4.1314	0.0000
171	0.0476	0.0365	1.3041	0.0961
173	-0.0067	0.0365	-0.1838	0.5729
174	0.1946	0.0365	5.3302	0.0000
190	0.3182	0.0365	8.7138	0.0000
191	-0.0169	0.0365	-0.4642	0.6787
240	0.3322	0.0365	9.0974	0.0000
241	0.1946	0.0365	5.3302	0.0000
260	-0.0033	0.0365	-0.0916	0.5365
270	0.3119	0.0365	8.5412	0.0000
271	0.2271	0.0365	6.2199	0.0000
290	0.1310	0.0365	3.5885	0.0002
291	0.4621	0.0365	12.6543	0.0000
Acc	0.3986	0.0365	10.9155	0.0000
Overall	0.2884	0.0164	17.5413	0.0000

Figure 11 Fleiss' Kappa statistics of the codes among the three auditors

Each auditor vs. experts, Fleiss' Kappa

This compares the two output from each auditor to that of the experts. There are several empty values because none of the auditors assigned those codes to any of the tickets.

Fleiss' Kappa Statistics

Appraiser	Response	Kappa	SE Kappa	Z	P(vs > 0)
1	140	*	*	*	*
	152	-0.0101	0.1000	-0.1010	0.5402
	170	-0.0490	0.1000	-0.4895	0.6878
	171	*	*	*	*
	173	*	*	*	*
	174	*	*	*	*
	190	0.4214	0.1000	4.2145	0.0000
	191	-0.0101	0.1000	-0.1010	0.5402
	240	0.3296	0.1000	3.2960	0.0005
	241	*	*	*	*
	260	*	*	*	*
	270	0.4345	0.1000	4.3450	0.0000
	271	*	*	*	*
	290	-0.0101	0.1000	-0.1010	0.5402
	291	0.6564	0.1000	6.5636	0.0000
	Acc	0.2385	0.1000	2.3853	0.0085
	Overall	0.2483	0.0481	5.1666	0.0000

Appraiser	Response	Kappa	SE Kappa	Z	P(vs > 0)
2	140	*	*	*	*
	152	-0.0101	0.1000	-0.1010	0.5402
	170	0.2591	0.1000	2.5906	0.0048
	171	-0.0257	0.1000	-0.2567	0.6013
	173	*	*	*	*
	174	*	*	*	*
	190	0.1980	0.1000	1.9805	0.0238
	191	*	*	*	*
	240	0.2498	0.1000	2.4875	0.0064
	241	*	*	*	*
	260	*	*	*	*
	270	0.2161	0.1000	2.1609	0.0154
	271	-0.0153	0.1000	-0.1525	0.5606
	290	-0.0257	0.1000	-0.2567	0.6013
	291	0.7396	0.1000	7.3958	0.0000
	Acc	0.5776	0.1000	5.7756	0.0000
	Overall	0.3490	0.0480	7.2683	0.0000

Appraiser	Response	Kappa	SE Kappa	Z	P(vs > 0)
3	140	*	*	*	*
	152	*	*	*	*
	170	0.1828	0.1000	1.8284	0.0337
	171	*	*	*	*
	173	*	*	*	*
	174	*	*	*	*
	190	0.7469	0.1000	7.4685	0.0000
	191	*	*	*	*
	240	0.1246	0.1000	1.2456	0.1065
	241	*	*	*	*
	260	*	*	*	*
	270	0.3688	0.1000	3.6885	0.0001
	271	*	*	*	*
	290	*	*	*	*
	291	0.3125	0.1000	3.1249	0.0009
	Acc	0.6886	0.1000	6.8865	0.0000
	Overall	0.4834	0.0503	9.6057	0.0000

* When all sample standards and responses of a trial(s) equal the value or none of them equals the value, kappa cannot be computed.

Each auditor vs. experts, Kendall's statistics (Risk Level Only)

This is the Kendall's correlation coefficients. The coefficient of the top represents the ordered correlation between the two output of auditor 1 (or 9), whereas the two coefficients of the bottom represents the ordered correlation between the two assessments of auditor 1 (9) to the experts'. Note that, unlike other examples listed on the Appendix, this example is generated from the scenario where all risk codes were grouped by risk levels ranging from 1 to 4, with higher number representing higher risk level. This is because the main advantage of the Kendall's coefficient, as opposed to a regular Pearson's and the Fleiss' Kappa, is that it adjusts its coefficients of ordinal variables by how "close" or "far" the two paired values are.

Kendall's Correlation Coefficient

Appraiser	Coef	SE Coef	Z	P
1	0.4132	0.0690	5.9813	0.0000

Kendall's Correlation Coefficient

Appraiser	Coef	SE Coef	Z	P
DPS09_10DEC2019	0.4699	0.0976	4.8071	0.0000
DPS09_31DEC2019	0.3564	0.0976	3.6434	0.0003

Figure 12 Kendall's correlation coefficients

All auditors vs. experts, assessment agreement

The figure below represents the assessment agreement among all auditors AND the experts. Out of 50 tickets inspected, four (8 % of all tickets) were unanimously agreed upon by the three auditors, as well as the experts, in both dates. Note that this percentage is smaller than that when assessment agreement was calculated only amongst the auditors. This is because in one ticket, the experts did not agree with the decision that was unanimous amongst auditors.

Assessment Agreement

# Inspected	# Matched	Percent	95% CI
50	4	8.00	(2.22, 19.23)

 # Matched: All appraisers' assessments agree with the known standard.

Figure 13 Assessment agreement of each auditor vs. experts

All auditors vs. experts, Fleiss' Kappa

Lastly, this is the Fleiss' Kappa of the three auditors, both times, against the experts. Kappa values never reach above 0.90, indicating poor agreement between the auditors and the experts.

Fleiss' Kappa Statistics

Response	Kappa	SE Kappa	Z	P(vs > 0)
140	*	*	*	*
152	*	*	*	*
170	0.130982	0.0577350	2.2687	0.0116
171	*	*	*	*
173	*	*	*	*
174	*	*	*	*
190	0.455449	0.0577350	7.8886	0.0000
191	*	*	*	*
240	0.234303	0.0577350	4.0582	0.0000
241	*	*	*	*
260	*	*	*	*
270	0.339812	0.0577350	5.8857	0.0000
271	*	*	*	*
290	*	*	*	*
291	0.569477	0.0577350	9.8636	0.0000
Acc	0.501578	0.0577350	8.6876	0.0000
Overall	0.360581	0.0281750	12.7979	0.0000

** When all sample standards and responses of a trial(s) equal the value or none of them equals the value, kappa cannot be computed.*

Figure 14 Fleiss' Kappa of the three auditors against the experts

Appendix B

Gradient Boosted Machine Model for Predicting Safety Violations

Gradient Boosted Machine Model for Predicting Safety Violations

PHMSA Grant 693JK31940021PSDP Final Report

The work documented in this report was funded by grant 693JK31940021PSDP from the Pipeline and Hazardous Materials Safety Administration in the United States Department of Transportation.

Jennifer H Van Mullekom, PhD Eric Bae, MS
VIRGINIA TECH | STATISTICAL APPLICATIONS & INNOVATIONS GROUP | SEPTEMBER 28, 2020

Gradient Boosted Machine Model for Predicting Safety Violations

EXECUTIVE SUMMARY

As more customers employ VA811's web ticket systems to request utility identification prior to excavation, detection of potential errors in tickets with high accuracy is essential. The web ticketing system contains safety level risks in a higher proportion than direct phone calls to VA811 representatives. VA811 contracted with SAIG to create a predictive model to improve the safety violation detection rate of web entry tickets. The resulting predictive model will be implemented as part of a new risk-based audit plan. If used to audit 100% of the tickets, the predictive model identifies approximately 2 times the amount of safety violations compared to the current random audit procedure without increasing the overall percentage of tickets audited. This report summarizes the analytics process and benefits. In addition, it provides recommendations on implementing the model in the current audit plan. Finally, the report details next steps in the journey to a self-updating AI model.

INTRODUCTION

In an effort to improve the overall safety of underground utility identification in Virginia as well as to more efficiently use resources, VA811 contracted the Virginia Tech Statistical Applications and Innovations Group (VT SAIG) to begin their journey into the world of artificial intelligence (AI). AI has been defined as the theory and development of computer systems able to perform tasks that normally require human intelligence, such as visual perception, speech recognition, decision making, and translation between languages (Lexico: Oxford English and Spanish Dictionary, Thesaurus, and Spanish to English Translator, 2020). Machine learning can be described as algorithms that use statistics to find patterns in massive amounts of data. Data includes numbers, words, images, clicks ---anything that can be digitally stored (Hao, 2018).

The first step in a journey toward AI is to establish the effectiveness of a machine learning model. VT SAIG collaborators have developed a gradient boosted machine model for deployment within VA811's audit process. This report details the project and will be divided into 7 sections including Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, Deployment, and Path Forward.

BUSINESS UNDERSTANDING

The VA811 audit process is overviewed in **Figure 1. Current VA811 Web ticket audit process.** Approximately 35% of web tickets are audited annually through random audit and quality control procedures. During the Initial Screening phase of the audit process, tickets are assigned an audit code as shown in Table 1. For the purposes of modeling, these codes were converted into a binary classification of Safety Violation = "Yes" and Safety Violation = "No". This conversion was supported by initial work

for this project as detailed in the report entitled "*VA811 Safety Risk Rating Auditor Agreement: Accuracy, Reproducibility, and Repeatability*". Note the conversions from three-digit audit codes to binary values were processed on the basis of propensity for damage risk to underground utilities and safety risk to workers. The table also indicates which codes exclude tickets from the modeling process because they were not tickets from the routine business process.

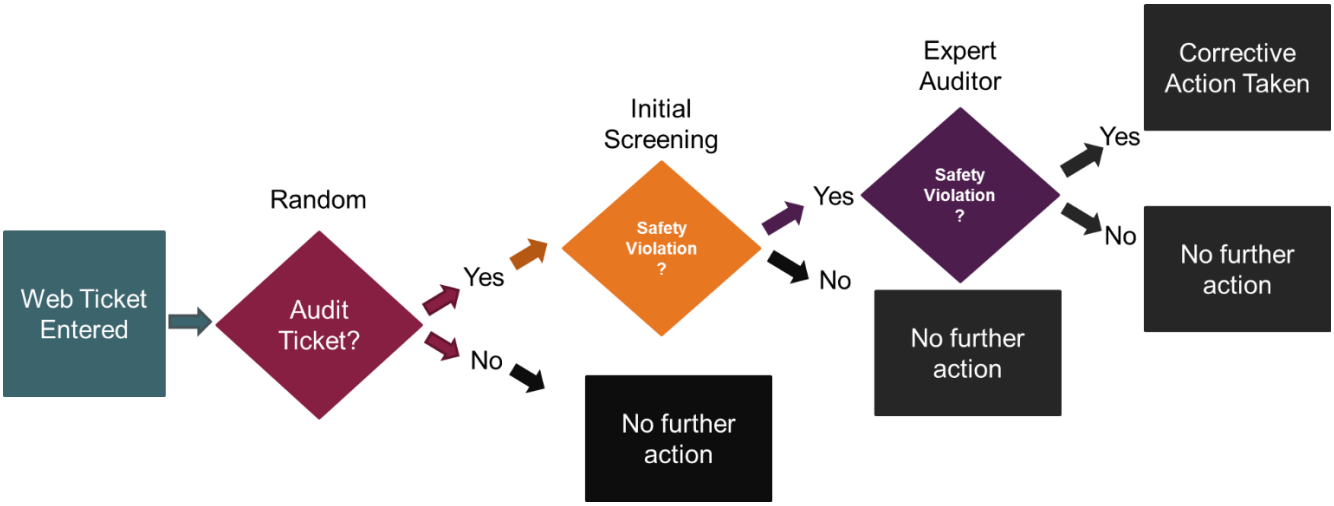


Figure 1. Current VA811 Web ticket audit process.

Table 1. VA811 audit codes and safety violation conversion.

Safety Level Violation in VA811 Comment Field Contents	Include in Modeling?	Safety Level Violation?
110 - Other was selected but details were not entered in excavation area field	Yes	No
140 - Street spelling or format does not match map	Yes	No
150 - Cross street(s) does not match map	Yes	No
151 - Street and cross street do not intersect	Yes	No
152 - No cross street - WTE Only	Yes	No
161 - 3HR Notice: over notification of utilities	Yes	No
162 - 3HR Notice: Instructions field is inaccurately formated	Yes	No
162 - 3HR Notice: Instructions field is inaccurately formatted	Yes	No
170 - Excavation area is vague - Ticket appears to be locatable	Yes	No
170 - Excavation area is vague. Ticket appears to be locatable	Yes	No
171 - For linear excavation, the excavation area exceeds one mile in length	Yes	No

Safety Level Violation in VA811 Comment Field Contents	Include in Modeling?	Safety Level Violation?
171 - Linear excavation does not include beginning/ending points (premarked)	Yes	No
172 - Driving directions are not entered (not a direct hit)	Yes	No
173 - Contains grammar or spelling errors	Yes	No
174 - Description of linear excavation exceeds one mile	Yes	No
181 - 3HR Update Remark:Contains data that refers to previously issued ticket	Yes	Yes
181 - Instructions field contains data that should have been removed: see guide	Yes	Yes
190 - Polygon does not cover description of excavation, utilities not missed	Yes	Yes
190 - Polygon does not match excavation area - utility(s) not missed	Yes	Yes
191 - Polygon was drawn in wrong area (utilities not missed)	Yes	Yes
192 - Excavation contains measurable distance not included in polygon: see guide	Yes	Yes
193 - SEG Tool used / polygon covered point data only	Yes	Yes
200 - 3HR:Improper use of filter - overnotification	Yes	Yes
201 - 3HR:Instructions field contains insufficient info or incorrectly formatted	Yes	Yes
240 - Incorrect address entered in Street field	Yes	Yes
241 - Incorrect street name entered in Street field	Yes	Yes
260 - Incorrect Ticket Type processed (Emergency or 3HR Notice)	Yes	Yes
261 -3HR Notice - under notification of utilities	Yes	Yes
262 - 3HR Notice: Mapping incorrect on original ticket	Yes	Yes
270 - Description of excavation not clear: see guide	Yes	Yes
270 - Specific location could be misinterpreted- may not be locatable	Yes	Yes
271 - Linear excavation does not include beginning/ending points (no premarks)	Yes	Yes
272 - Driving directions inaccurate	Yes	Yes
273 - Incorrect address(s) entered in Excavation Area field	Yes	Yes
290 - Polygon does not cover description of excavation - utilities were missed	Yes	Yes
290 - Polygon does not match excavation area - utility(s) missed	Yes	Yes
291 - Polygon was drawn in wrong area - utilities missed	Yes	Yes
300 - 3HR Notice - Improper use of filter - undernotification of utility(s)	Yes	Yes
Acceptable	Yes	No
Accurate	Yes	No
null	No?	Exclude
Performance Error-TL Use Only	No?	Exclude
Training Opportunity	No?	Exclude

**Note that some codes are listed twice because descriptions were amended in the ticket management system during the 2017-2019 time period. All code descriptions are included for completeness.*

VA811 processes approximately 600K web entry tickets per year with approximately 3% of those resulting in a safety violation classification after the expert audit phase of the process.

DATA PREPARATION

A substantial portion of any modeling project is spent in data preparation. This includes data exploration, data cleaning, and feature engineering. Data exploration includes univariate visualization and descriptive statistics that provide insights into data quality. In addition, this phase often includes simple fitting of the dependent variable (safety violation) versus the various independent variables taken one at a time. Based on the results of data exploration, it is often necessary to clean data in order to prepare data for modeling. Examples of data cleaning include making decisions about missing values, correcting values that are out of range, and collapsing categories. At this time the modeling team often makes decisions about what variables to include or not include in the model building phase of a data science project. Finally, features are engineered to create new candidate independent variables to help improve the prediction quality of the models. Engineered features are variables created from existing data based on subject matter expertise which capture additional information not represented in typical data collection.

VA811 DATA OVERVIEW

VA811 provided data which included six ticket files from the basic business process from years 2017, 2018, and 2019. The ticket file names are listed below:

- Ticket_Data2017p1.csv
- Ticket_Data2017p2. csv
- Ticket_Data2018p1. csv
- Ticket_Data2018p2. csv
- Ticket_Data2019p1. csv
- Ticket_Data2019p2. csv

An audit data file, named Verify_Data_040320.xlsx, included the results of the random ticket audits from 2017-2019 with the final audit code representing the expert audit decision from Figure 1. A list of the fields from each of these files along with their data definitions is listed in **Table 2. List and Description of Ticket Data Variables** and **Table 3. List and Description from Verify Data (Audited Ticket Data)**.

Table 2. List and Description of Ticket Data Variables

Field	Data Type	Example	Notes
id	Long Integer	1	Auto-assigned id number
ticket	Short Text	A123456789	Identification number assigned to all tickets
revision	Short Text	00A	Revision code at the end of a ticket number
original_ticket	Short Text	A123456789	Identification number for original ticket - issued prior to current ticket
original_date	Date With Time	9/28/2016 11:00	Date and time original ticket was submitted
original_account	Short Text	WJSMITH	Username for account that created the original ticket
replaced_by_ticket	Short Text	-	Defaults to empty, not used by system or VA811
replace_by_date	Date With Time	9/28/2016 11:00	Date and time ticket must be updated by
reference	Short Text	05-10194055	References existing Work Order Number found on a ticket that is being revised
account	Short Text	WJSMITH	Account username
channel	Short Text	WEB	Channel ticket was entered through
taken_source	Short Text	H5TE INHSE	Software used to created ticket
taken_version	Short Text	1.0.33	Software version
started	Date With Time	9/28/2016 11:00	Date and time ticket was started
completed	Date With Time	9/28/2016 11:00	Date and time ticket was submitted
type	Short Text	UPDT	Type of Ticket generated - NEW, UPDT, RMRK, 3HRS, CNCL
priority	Short Text	RUSH	Delivery requirement based on ticket type
category	Short Text	LREQ	Defaults to LREQ unless it's a 911 generated ticket
lookup	Short Text	STRT	Unknown - but no impact on the VT project
caller_type	Short Text	OWNR	Type of Caller
name	Short Text	SOME COMPANY	Name of Caller or Business Name
address1	Short Text	8132 Lee Hwy	Address Number and Street Name of the Caller
address2	Short Text	-	Not used by VA811
city	Short Text	Falls Church	Name of Town/City where the Caller is located
cstate	Short Text	VA	State in which the Caller is located
zip	Short Text	22042	Zip code in which the Caller is located
phone	Short Text	7035606222	Main phone number of Caller or Business
phone_ext	Short Text	111	Phone extension

Field	Data Type	Example	Notes
caller	Short Text	JOHN SMITH	Name of the Caller
caller_phone	Short Text	7035606222	Same or alternate phone number in which to reach the Caller
caller_phone_ext	Short Text	111	Phone extension to same or alternate phone number in which to reach the Caller
contact	Short Text	JOHN SMITH	Additional Contact aka Field Contact
contact_phone	Short Text	7039297021	Phone number for additional contact
contact_phone_ext	Short Text	111	Phone extension for additional contact
fax	Short Text	8775676787	Fax Number in which to receive ticket and PRS confirmations
fax_ext	Short Text	111	Phone extension for fax number
pager	Short Text	-	Not used by VA811
pager_ext	Short Text	-	Not used by VA811
cell	Short Text	-	Not used by VA811
cell_ext	Short Text	-	Not used by VA811
email	Short Text	johnsmith@test.com	Email address in which to receive ticket and PRS confirmations
best_time	Short Text	-	Not used by VA811
st	Short Text	51	First 2 digits within a FIPS code that identifies the State (51 aka Virginia)
co	Short Text	177	Last 3 digits within a FIPS code that identifies the county/city (177 aka Spotsylvania)
fips	Short Text	95291	A Five Digit code that uniquely identifies Virginia Places (city, town village), Urbanized Areas, Urban Clusters, Micropolitan/Metropolitan Statistical Areas
map	Short Text	51177	Federal Information Processing Standard (FIPS) A Five Digit code that uniquely identifies counties and county equivalents (cities within a commonwealth)
state	Short Text	VA	The State in which Excavation will be taking place (hard coded in TE)
county	Short Text	Spotsylvania	County or City identified as to where the Excavation is taking place
place	Short Text	Berkeley	Places (city, town village), Urbanized Areas, Urban Clusters, Micropolitan/Metropolitan Statistical Areas as identified within VA
inside_outside	Short Text	B	Not used by VA811
subdivision	Short Text	Arcadia Crossing - South	Known subdivision or Business Name
lot	Short Text	1A	Lot number assigned to property
st_from_address	Short Text	100	Beginning Address Number
st_to_address	Short Text	100	Ending Address Number, can be the same as the st_from_address or incremental
street	Short Text	N Toano Dr SW	Name of Street
cross1	Short Text	New Market Ct	Name of Cross Street 1
cross2	Short Text	Solitaire Ln	Name of Cross Street 2
st_prefix	Short Text	N	Prefix of Street within Street field

Field	Data Type	Example	Notes
st_name	Short Text	Toano	Street Name within Street field
st_type	Short Text	Dr	Street Type within Street field
st_suffix	Short Text	SW	Street Suffix within Street field
st_recno	Long Integer	0	Not used by VA811
c1_prefix	Short Text	N	Prefix of Street within Cross St 1 field
c1_name	Short Text	New Market	Street Name within Cross St 1 field
c1_type	Short Text	Ct	Street Type within Cross St 1 field
c1_suffix	Short Text	S	Street Suffix within Cross St 1 field
c1_recno	Long Integer	0	Not used by VA811
c2_prefix	Short Text	S	Prefix of Street within Street field
c2_name	Short Text	Solitaire	Street Name within Cross St 2 field
c2_type	Short Text	Ln	Street Type within Cross St 2 field
c2_suffix	Short Text	S	Street Suffix within Cross St 2 field
c2_recno	Long Integer	0	Not used by VA811
latitude	Short Text	37.29431	Latitude Coordinates provided by the caller
longitude	Short Text	-77.30706	Longitude coordinates provided by the caller
side_of_street	Short Text	-	Not used by VA811
side_of_lot	Short Text	-	Not used by VA811
work_date	Date With Time	1/3/2020 7:00:00 AM	Date and Time that all Utilities should have responded by and Excavation can begin
meet_date	Date With Time	1/6/2020 11:00:00 AM	Date and Time provided by the caller for when the Meet should occur
response_due	Date With Time	1/3/2020 7:00:00 AM	Date and Time by which the Utilities are to respond to the ticket
project_end_date	Date With Time	1/1/1900 or 1/1/2000	Default date generated by the system
expires	Date With Time	1/23/2020 7:00:00 AM	Date and Time that the ticket expires
hours_notice_clock	Integer	141	Actual Hours between ticket completion and legal work date
hours_notice_business	Integer	72	Hours between ticket completion and legal work date based on working days as described in the DPA
work_type	Short Text	GAS MAIN - REPAIR, REPLACE OR ABANDON	Type of work taking place
duration	Short Text	-	Not used by VA811
done_for	Short Text	Verizon	Who the work is being done for
header	Short Text	-	Not used by VA811

Field	Data Type	Example	Notes
uob	Short Text	U	Defaults to U (underground)
from_rr_marker	Short Text	-	Not used by VA811 - from railroad marker
to_rr_marker	Short Text	-	Not used by VA811 - to railroad marker
rr_subdivision	Short Text	-	Not used by VA811 - railroad subdivision
permit_number	Short Text	W/O#: 3445249 Permit #: TES2019-04057	Work Order and/or Permit #, neither is required
license_no	Short Text	875	A designer's professional license number, required for Designer tickets
map_reference	Short Text	267J5	ADC grid and number from map layers that is provided on the ticket output for Members to help identify location/aid in routing of tickets (where available)
extent_top	Decimal	36.745933	Furthest top point of notification polygon
extent_left	Decimal	-76.02721	Furthest left point of notification polygon
extent_bottom	Decimal	36.744603	Furthest bottom point of notification polygon
extent_right	Decimal	-76.025168	Furthest right point of notification polygon
bestfit_y1	Decimal	36.744557	Furthest extents of notification polygon as lat/long coordinate (simple re-creation)
bestfit_x1	Decimal	-76.027082	Furthest extents of notification polygon as lat/long coordinate (simple re-creation)
bestfit_y2	Decimal	36.745766	Furthest extents of notification polygon as lat/long coordinate (simple re-creation)
bestfit_x2	Decimal	-76.027327	Furthest extents of notification polygon as lat/long coordinate (simple re-creation)
bestfit_y3	Decimal	36.746163	Furthest extents of notification polygon as lat/long coordinate (simple re-creation)
bestfit_x3	Decimal	-76.025371	Furthest extents of notification polygon as lat/long coordinate (simple re-creation)
bestfit_y4	Decimal	36.744954	Furthest extents of notification polygon as lat/long coordinate (simple re-creation)
bestfit_x4	Decimal	-76.025126	Furthest extents of notification polygon as lat/long coordinate (simple re-creation)
centroid_y	Decimal	38.889071	Center of notification polygon as latitude coordinate
centroid_x	Decimal	-77.102362	Center of notification polygon as longitude coordinate
area_in_miles	Decimal	0.064441	Total notification area in miles
intersection	Yes/No	-	Not used by VA811
blasting	Yes/No	-1 or 0	Will there be blasting, yes or no
boring	Yes/No	-1 or 0	Will there be boring (horizontal directional drilling), yes or no
rr	Yes/No	-	Not used by VA811
emergency	Yes/No	-1 or 0	Emergency ticket type
plan_design	Yes/No	-1 or 0	Designer ticket type
meet	Yes/No	-1 or 0	Meet ticket type
special_project	Yes/No	-1 or 0	Special Project ticket type

Field	Data Type	Example	Notes
right_of_way	Yes/No	-	Not used by VA811
permit	Yes/No	-	Not used by VA811
no_queue	Yes/No	-1 or 0	Do not queue ticket upon inquiry (ticket has been blocked)
checked_map	Yes/No	-1 or 0	Default to true, map is always checked (Member Lookup)
grids_from_map	Yes/No	-1 or 0	Defaults to true, pulls grids from map that notification polygon intersects and shows grid names under Full Ticket Display and on ticket output
white_paint	Yes/No	-1 or 0	Is the site pre-marked
response_required	Yes/No	-1 or 0	Is a Response Required based on ticket type and priority
first_time_caller	Yes/No	-1 or 0	First Time Caller
street_not_in_map	Yes/No	-1 or 0	Checkbox used to indicate a discrepancy within map data
spare_bit	Yes/No	-1 or 0	Not Used
location	Long Text	Entire Property	Excavation Area
remarks	Long Text	Reason Cancelled: Incorrect Mapping	Instruction Field
comments	Long Text	New Street	In-House Comments, not transmitted on ticket output

Table 3. List and Description from Verify Data (Audited Ticket Data)

Field	Data Type	Example	Notes
id	Long Integer	1	Auto-assigned id number
ticket	Short Text	A123456789	Identification number assigned to all tickets
revision	Short Text	00A	Revision code at the end of a ticket number
ticket_completed	Date With Time	9/28/2016 11:00	Date and time ticket was submitted
ticket_operator	Short Text	WJSMITH	Username for account that created the ticket
ticket_operator_type	Short Text	C	Type of operator, indicated as 'C' for customer service rep or 'R' for remote user
ticket_priority	Short Text	RUSH	Delivery requirement based on ticket type
ticket_type	Short Text	UPDT	Type of Ticket generated - NEW, UPDT, RMRK, 3HRS, CNCL
verified_date	Date With Time	9/28/2016 11:00	Date the audit was performed
verifier	Short Text	1ABC	Username for account that audited the ticket
verify	Short Text	G	Designation for type of audit (G, B, U, F, C)
comment	Short Text	290 - Polygon does not match excavation area - utility(s) missed	Audit code
comment_free	Long Text	PHONE: BLOCK, called the address in the excavation does not match address listed	Text field for auditor's initial comments. Used for follow up comments after audit is reviewed
review	Yes/No	-1	Field to mark audit as reviewed or not reviewed
reviewer	Short Text	1ABC	Username for account that reviewed the audit
reviewed	Date With Time	9/28/2016 11:00	Date the audit was marked as reviewed
changed	Date With Time	9/28/2016 11:00	Date audit was changed (ticket was re-audited). Field remains null on latest audit

DATA EXPLORATION

VT SAIG has created an R program that analyzes categorical, count, and continuous independent variables by providing descriptive statistics and basic visualizations. This report creates an indexed *.HTML file with an individual report on each variable. For reference, (R Statistical Programming Language, 2020) is an open source statistical programming language. Examples of output from this file are shown in **Figure 2**, **Figure 3**, and **Figure 4**.

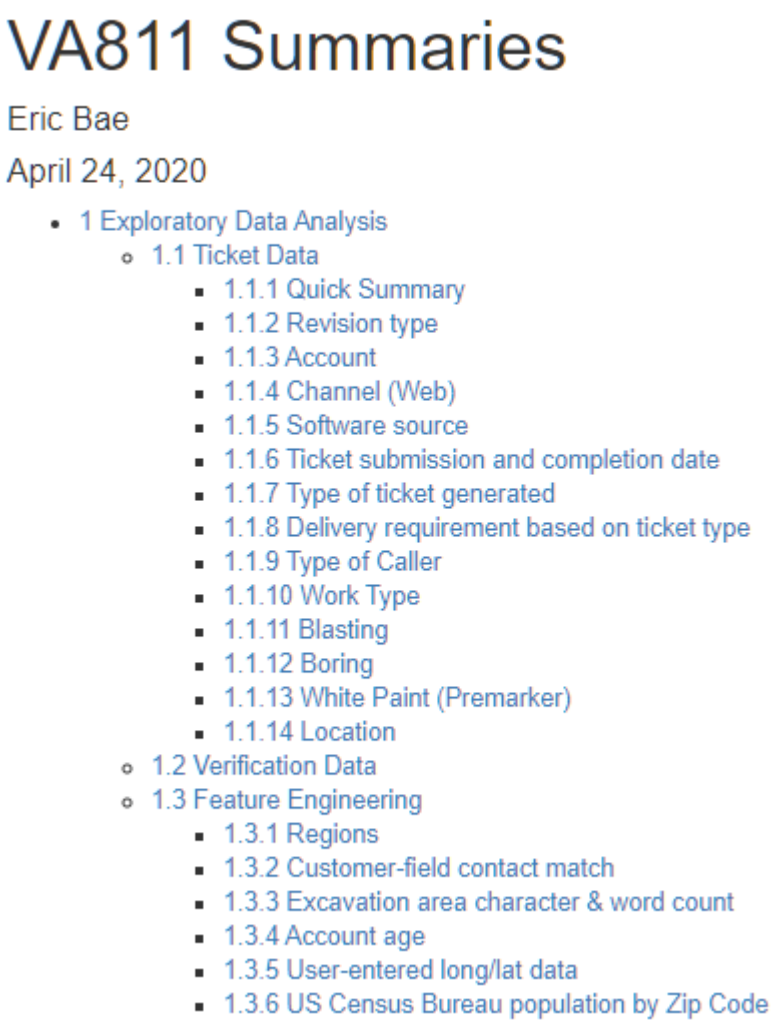


Figure 2. Example of Table of Contents from .HTML Data Exploration Report Generated in R

1.1.13 White Paint (Premarker)

Premarked	Frequency
FALSE	1318172
TRUE	380007

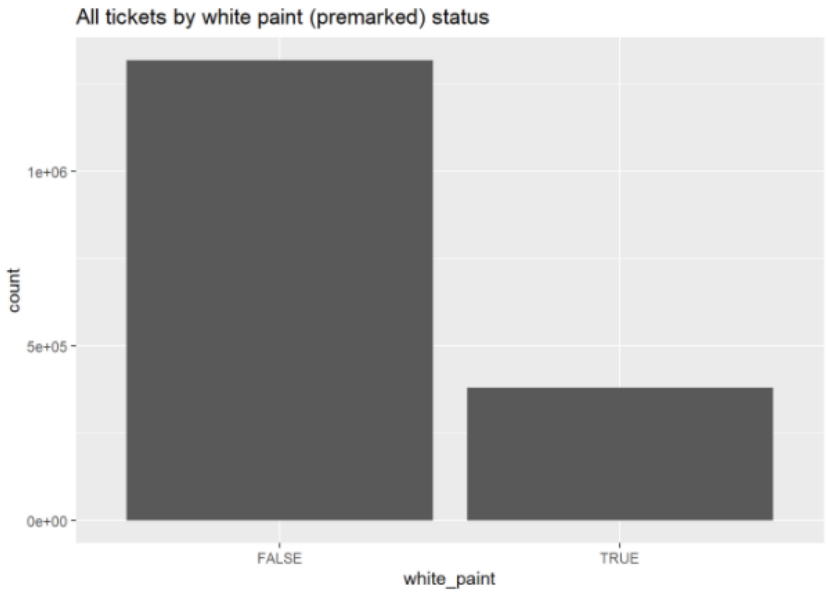
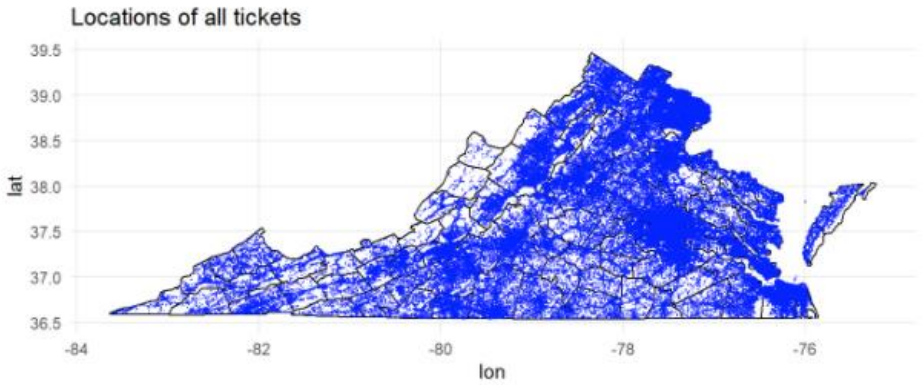


Figure 3. Example of Variable Frequency Summary from .HTML Data Exploration Report Generated in R

1.1.14 Location

There were 4 tickets without location information in addition to the nine already mentioned above. Those are ticket numbers B701100144, B704700130, B703101828, B706500341.



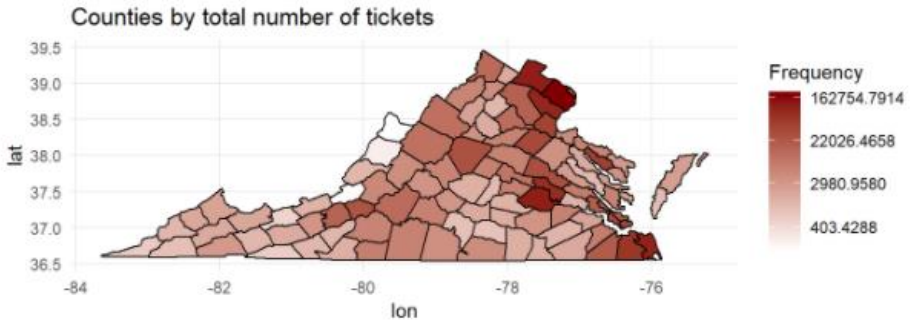


Figure 4. Example of Map Ticket Summaries from *.HTML Data Exploration Report Generated in R

Multiple iterations of this report were generated and discussed with VA811 during the data exploration process as data was cleaned and prepared for modeling.

DATA CLEANING

Administrative data from data systems designed for running business processes contain data that is perfectly suited for its intended purpose. However, this data is not necessarily formatted for statistical modeling. Of course, even in the most closely curated data systems, some data quality issues also need to be addressed.

Based on the data exploration phase, examples of data cleaning resolving merge non-matches, collapsing categories, etc. were carried out iteratively with the VA811 team. A complete list of data cleaning as well as the R code which cleans the data is shown in **Table 4. Data Cleaning Issues and R Code Calculations**.

Table 4. Data Cleaning Issues and R Code Calculations

Cleaned Data	Calculation
Remove 21 tickets	All 2016 tickets and tickets who do not appear in main data but in verify R code: date_start <- as.Date("01/01/2017", format = "%m/%d/%Y") Verify_Data_New <- Verify_Data_New[!(Verify_Data_New$ticket_completed < date_start),]
Caller Type	ENGR and OWNR merged into CONT R code: # Merging Owner & Engineers to Contractors Data_All_New$caller_type[which(Data_All_New$caller_type %in% c("ENGR", "OWNR"))] <- "CONT" levels(Data_All_New$caller_type) <- c("CONT", "CONT", "CONT", "UTIL")
Types of Tickets	Tickets whose types are “CNCL” meaning canceled were removed. There are so many reasons they are canceled and for auditing purposes they are useless. However, they could be looked into for other reasons. R code: index_type_cancel <- which(Data_All_New$type == "CNCL") Data_All_New[-index_type_cancel,]

Total number of seconds tickets were open	All tickets whose numbers > 3,600 seconds (30 minutes) adjusted to 3,600. This is due to the knowledge that system is supposed to time out after 30 minutes; anything above that is not possible.
	
	R code:
	
	# Ticket processing time and age
	Data_ticket_time <- difftime(Data_All_New$completed,
	Data_All_New$started,
	units = "auto")
	Data_ticket_age <- difftime(max(Data_All_New$completed),
	Data_All_New$started,
	units = "auto")
	
	started_index <- which(is.na(Data_All_New$started))
	completed_index <- which(is.na(Data_All_New$completed))
	Data_ticket_time[unique(c(started_index, completed_index))] <- NA
	Data_ticket_age[unique(c(started_index, completed_index))] <- NA

Cleaned Data	Calculation
	Data_ticket_time[which(Data_ticket_time > 60*60)] <- 60*60 Data_ticket_time <- data.frame(ticket_time = Data_ticket_time, ticket_age = Data_ticket_age) Data_All_New$ticket_process_time <- Data_ticket_time$ticket_time Data_All_New$ticket_age <- Data_ticket_time$ticket_age

FEATURE ENGINEERING

As mentioned previously, feature engineering often improves the predictive quality of machine learning models by capturing variability not represented in existing data fields. The features in Table 5 were engineered from the Ticket Data and Verify files provided by VA811. Several features are simply calculations like the elapsed time to complete a ticket entry based on the time the ticket started and ended. Some of these features represent summary statistics about accounts and web ticket locations such as the number of tickets entered per account in the last 24 hours. Other features use natural language processing of the excavation location description in order to identify phrases and words likely to be associated with a safety violation.

For the latter, JMP® (JMP® Version 15, 1989-2020) software was used to identify words and phrases associated with a safety violation via Natural Language Processing (NLP) techniques in the Text Explorer platform. Thousands of words and phrases were identified as potentially predictive in this process. Notable findings from this analysis are the phenomena of ticket entry operators "copying and pasting" location descriptions over multiple tickets and the fact that the use of direction words (North, South, East, West, etc.) are associated with safety violations. Preliminary analyses were executed to reduce the number of words and phrases input into the predictive model. The words and phrases ultimately identified for input into the model were engineered in R. A description of the engineered features and their accompanying R code is found in ***Table 5. Description of Engineered Features including R Code Calculations.***

Table 5. Description of Engineered Features including R Code Calculations.

Engineered Feature	Source	Unit	Calculation

Safety Violation (Response)	Verify data – comment	1 – Yes
		0 – No
			The first three digits of Comment <=180 or "Acc" is "No" violation. The first three digits of Comment>180 is "Yes". Everything else has been assigned "Ignore."
			
			R code:
			
			# Violation status
			verify_comment <- substring(Verify_Data_New$comment, 1, 3)
			
			verify_exclude <- which(verify_comment %in% c("nul", "Per", "Tra"))
			verify_index_safe <- which(verify_comment %in%
			c("Acc", "110", "140", "150", "151",
			"152", "161", "162", "170",
			"171", "172", "173", "174"))
			
			verify_comment[-c(verify_index_safe, verify_exclude)] <- "Yes"
			verify_comment[verify_index_safe] <- "No"
			verify_comment[verify_exclude] <- "Ignore"
			Verify_Data_New$violation_stat <- verify_comment

Not Violation	Main data – account Verify data – comment	Integers (0, 1, …)	Total number of accounts that were verified and were found to be not in violation of the safety code (either no violation or has violation codes with minimal risk – those in 170s or less) R Code: violation_number <- as.data.frame.matrix(table(Data_Merged$account, Data_Merged$violation_stat)) violation_number <- cbind.data.frame(account = rownames(violation_number), violation_number) colnames(violation_number) <- c("account", "ignore", "not_violation", "violation") violation_number <- violation_number[, -2] Data_All_New_New <- merge(Data_All_New, violation_number, by = "account")

Table 5. Description of Engineered Features including R Code Calculations.

Engineered Feature	Source	Unit	Calculation
Violation	Main data – account Verify data - comment	Integers (0, 1, …)	Total number of accounts that were verified and found to be IN violation of the safety code (whose violation codes are 180 and above) R Code: See above

Total number of seconds tickets were open	Main data – ticket submission & completion date	Seconds	Find the differences (in seconds) between the ticket completion to ticket submission dates.
			
			R code:
			
			# Ticket processing time and age
			Data_ticket_time <- difftime(Data_All_New$completed,
			Data_All_New$started,
			units = "auto")
			Data_ticket_age <- difftime(max(Data_All_New$completed),
			Data_All_New$started,
			units = "auto")
			
			started_index <- which(is.na(Data_All_New$started))
			completed_index <- which(is.na(Data_All_New$completed))
			Data_ticket_time[unique(c(started_index, completed_index))] <- NA
			Data_ticket_age[unique(c(started_index, completed_index))] <- NA
			Data_ticket_time[which(Data_ticket_time > 60*60)] <- 60*60
			Data_ticket_time <- data.frame(ticket_time = Data_ticket_time,
			ticket_age = Data_ticket_age)
			
			Data_All_New$ticket_process_time <- Data_ticket_time$ticket_time
			Data_All_New$ticket_age <- Data_ticket_time$ticket_age

Work category	Main data – work type	24 categories (e.g. Communications, Gas)	There were 138 work types. Since this was too many to analyze, they were grouped and reduced down to 24 work categories. R code: table_work_category <- table(Data_Merged$Category)

Table 5. Description of Engineered Features including R Code Calculations.

Engineered Feature	Source	Unit	Calculation
			table_work_category <- sort(table_work_category, decreasing = TRUE) work_category_index <- which(is.na(Data_Merged$Category)) M_missing_work_category <- Data_Merged[work_category_index,] kable(table_work_category, col.names = c("Work Category", "Frequency")) Data_Merged$Category <- factor(Data_Merged$Category, levels = names(table_work_category))

Regions	Main data – county	Central VA
		Eastern VA
		Fredericksburg Area
		Lynchburg/Danville Area
		Northern VA
		Southwestern VA
		Western VA
		Winchester Area
			All counties/cities were grouped into 8 different regions.
			
			<u>R code</u>:
			
			# Regions data
			Regions <- as.vector(read_excel("DPC Regions April 2020.xlsx"))
			Regions_SWVA <- toupper(Regions$`Southwestern VA`)
			Regions_WVA <- toupper(Regions$`Western VA`)
			Regions_FA <- toupper(Regions$`Fredericksburg Area`)
			Regions_NVA <- toupper(Regions$`Northern VA`)
			Regions_EVA <- toupper(Regions$`Eastern VA`)
			Regions_WA <- toupper(Regions$`Winchester Area`)
			Regions_CVA <- toupper(Regions$`Central VA`)
			Regions_LDA <- toupper(Regions$`Lynchburg/Danville Area`)
			
			Data_All_New$region <- rep(NA, nrow(Data_All_New))
			Data_All_New$region[which(Data_All_New$county %in% Regions_SWVA)] <-
			"Southwestern VA"
			Data_All_New$region[which(Data_All_New$county %in% Regions_WVA)] <-
			"Western VA"
			Data_All_New$region[which(Data_All_New$county %in% Regions_FA)] <-
			"Fredericksburg Area"
			Data_All_New$region[which(Data_All_New$county %in% Regions_NVA)] <-
			"Northern VA"
			Data_All_New$region[which(Data_All_New$county %in% Regions_EVA)] <-
			"Eastern VA"
			Data_All_New$region[which(Data_All_New$county %in% Regions_WA)] <-

Table 5. Description of Engineered Features including R Code Calculations.

Engineered Feature	Source	Unit	Calculation
			"Winchester Area" Data_All_New$region[which(Data_All_New$county %in% Regions_CVA)] <- "Central VA" Data_All_New$region[which(Data_All_New$county %in% Regions_LDA)] <- "Lynchburg/Danville Area"
Excavation Area by Number of Characters	Main data – location	Integers (0, 1, …)	Counted the total number of characters in the variable "location." R code: # Adding Number of Characters and Number of Words in Location Variable as Variables Data_All_New$number_characters <- nchar(gsub(" ", "", Data_All_New$location)) remove_spaces <- gsub(" ", "", Data_All_New$location) remove_spaces <- gsub("\n", "", remove_spaces) remove_spaces <- gsub(" ", " ", remove_spaces) Data_All_New$number_words <- str_count(remove_spaces, " ") + 1
Excavation Area by Number of Words	Main data – location	Integers (0, 1, …)	Counted the number of spaces in the variable "location" and added 1. R code: # Adding Number of Characters and Number of Words in Location Variable as Variables Data_All_New$number_characters <- nchar(gsub(" ", "", Data_All_New$location)) remove_spaces <- gsub(" ", "", Data_All_New$location) remove_spaces <- gsub("\n", "", remove_spaces) remove_spaces <- gsub(" ", " ", remove_spaces) Data_All_New$number_words <- str_count(remove_spaces, " ") + 1

Account Age	Main data – completed date	Seconds (≥ 0)	Difference between the completed date of each ticket from the final day of the data (31-Dec-2019).
			
			R code:
			
			# Ticket processing time and age
			Data_ticket_time <- difftime(Data_All_New$completed,
			Data_All_New$started,

Table 5. Description of Engineered Features including R Code Calculations.

Engineered Feature	Source	Unit	Calculation
			units = "auto") Data_ticket_age <- difftime(max(Data_All_New$completed), Data_All_New$started, units = "auto") started_index <- which(is.na(Data_All_New$started)) completed_index <- which(is.na(Data_All_New$completed)) Data_ticket_time[unique(c(started_index, completed_index))] <- NA Data_ticket_age[unique(c(started_index, completed_index))] <- NA Data_ticket_time[which(Data_ticket_time > 60*60)] <- 60*60 Data_ticket_time <- data.frame(ticket_time = Data_ticket_time, ticket_age = Data_ticket_age) Data_All_New$ticket_process_time <- Data_ticket_time$ticket_time Data_All_New$ticket_age <- Data_ticket_time$ticket_age
Months	Main data – completed date	Factored values between 1 and 12	Tickets assigned by the month in which they were completed R code: Data_All_New$month <- as.factor(format(Data_All_New$completed, "%m"))
Time of day	Main data – completed date	Factored values between 0 and 23	Tickets assigned by the hour of day in which they were completed. (I.e. a ticket that was completed at 4:43 p.m. ET would have been assigned 16) R code: Data_All_New$time_of_day <- as.factor(format(Data_All_New$completed, "%H"))
Violation rate by month	Number of tickets verified by month*, Safety Violation (Yes or No)*	Real value between approx. 0.019 to 0.031.	Found the number of safety violations by month, divided by the number of tickets verified by month. R code: violation_by_month <- table(Data_Merged_by_Verify$violation_stat, Data_Merged_by_Verify$month)

Table 5. Description of Engineered Features including R Code Calculations.

Engineered Feature	Source	Unit	Calculation
			violation_rate_by_month <- cbind.data.frame(Month = names(violation_by_month[3,]/ rowSums(violation_by_month[2:3,])), `Violation rate` = violation_by_month[3,]/ colSums(violation_by_month[2:3,]))
Directional Words	Main data – location	1 – Yes 0 – No	All tickets whose location input by caller contains any directional words, such as “South”, “West”, “East”, “North”, “Southwest”, “Southeast”, “Northwest”, “Northeast”, as well as their initials. R code: # Does the location comment contain directional words? Data_All_New$direction_words <- str_detect(Data_All_New$location, c(" E ", " S ", " W ", " N ", " SE ", " SW ", " NE ", " NW ", "EAST", "SOUTH", "WEST", "NORTH"))
Number of tickets with same or close match text by ticket operator	Main data – caller & contract	1 – close match 0 – not match	Calculated Levenshtein distance between the caller and contract input for every single ticket, then tickets whose values are higher than 0.5 were assigned 1 (close match) and others were assigned 0 (not match). R Code: caller <- Data_All_New$caller contact <- Data_All_New$contact Data_All_New$name_match <- ifelse(as.character(caller) == as.character(contact), 1, 0) caller[is.na(caller)] <- "" contact[is.na(contact)] <- "" caller_length <- str_length(as.character(caller)) contact_length <- str_length(as.character(contact)) name_length <- pmax(caller_length, contact_length)

Table 5. Description of Engineered Features including R Code Calculations.

Engineered Feature	Source	Unit	Calculation
			not_empty <- which(!is.na(contact)) empty <- which(is.na(contact)) match_rate <- sapply(not_empty, function(x) 1 - as.numeric(StrDist(as.character(caller[x]), as.character(contact[x])))/name_length[x]) #rm(list=setdiff(ls(), "match_rate")) Data_All_New$match_rate <- Data_All_New$close_name <- rep(0, nrow(Data_All_New)) Data_All_New$match_rate[not_empty] <- match_rate Data_All_New$match_rate[empty] <- Data_All_New$close_name[empty] <- NA Data_All_New$close_name[not_empty] <- ifelse(match_rate > 0.5, 1, 0)
Time of day ticket entered	Main data – completed date	Values between 00 to 23, with 00 representing midnight and 23 representing 11:00 p.m.	Looked at which hour of the day ticket was entered. R Code: Data_All_New$time_of_day <- as.factor(format(Data_All_New$completed, "%H"))

Number of tickets entered within 24 hours	Main data – completed date	Integer (0, 1, …)	For each caller-ticket combination, this feature looks at how many
			
			R Code:
			
			account_list <- unique(Verify_Merged_by_Data$account) # unique list of accounts
			verified_index <- which(!is.na(Verify_Merged_by_Data$verified_date)) # indices for verified tickets
			total_within_radius <- rep(NA, nrow(Verify_Merged_by_Data))
			total_ticket_num <- rep(NA, nrow(Verify_Merged_by_Data))
			Data_completed_date <- Verify_Merged_by_Data$completed +
			rnorm(nrow(Verify_Merged_by_Data), sd = 0.5) # Tiebreaker
			
			for (i in 1:length(account_list)) { # per account
			account_index <- which(Verify_Merged_by_Data$account == account_list[i]) # all tickets for each account
			comp_date <- Data_completed_date[account_index] # extract completed dates for that account
			comp_date_24hours <- cbind(comp_date - 3600*24, comp_date + 3600*24) # extract all tickets within 24 hour window
			for (j in 1:length(comp_date)) { # per ticket within 24 hour window, same account
			within_24hrs_index <- which((comp_date > comp_date_24hours[j, 1] &

Table 5. Description of Engineered Features including R Code Calculations.

Engineered Feature	Source	Unit	Calculation

			comp_date < comp_date_24hours[j, 2]))
			exact_index <- which(comp_date == comp_date[j])
			Data_sub <- Verify_Merged_by_Data[within_24hrs_index,]
			Data_sub_sub <- cbind(Data_sub$centroid_x, Data_sub$centroid_y) # subsample data of only the tickets selected (24 hour window, same account)
			within_radius_sub <- matrix(NA, nrow(Data_sub_sub), nrow(Data_sub_sub)) # design matrix whose entries are binary 1 and 0
			total_ticket_num[account_index[j]] <- length(within_24hrs_index) - 1
			if (account_index[j] %in% verified_index) {
			for (k in 1:nrow(Data_sub_sub)) { # row of design matrix
			for (l in 1:nrow(Data_sub_sub)) { # column of design matrix
			within_radius_sub[k, l] <- ifelse(sqrt(sum((Data_sub_sub[k,] - Data_sub_sub[l,])^2)) < 1000/(60*5280), 1, 0) # whether the ticket is within 1,000 ft of each other or not
			}
			}
			total_within_radius[account_index[j]] <-
			rowSums(within_radius_sub)[which(within_24hrs_index == exact_index)] - 1 # total number of tickets within 1000 ft radius, 24 hour window AND is verified
			}
			}
			}
			
			Verify_Merged_by_Data$total_within_radius_24hrs <- total_within_radius # Add the variable to the grand matrix
			Verify_Merged_by_Data$total_ticket_num_24hrs <- total_ticket_num # Add the variable to the grand matrix

Number of tickets entered within a distance	Main data – centroid.x & centroid.y	Integer (0, 1, …)	Uses the same data as above, but also calculates how many other tickets were within a 1,000 feet radius of each ticket. R Code: See above (Number of tickets entered within 24 hours)

Table 5. Description of Engineered Features including R Code Calculations.

Engineered Feature	Source	Unit	Calculation

Correct state names	Main data – location	1 – Yes
		0 – No
			If the ticket location text input includes the exact address that is to be serviced, returns 1. Otherwise, 0.
			
			R Code:
			
			# Add street name info
			street_names <- with(Data_Merged_by_Verify,
			ifelse(st_from_address == 0,
			paste(street), paste(st_from_address, street)))
			street_names <- gsub("\\", "", street_names, fixed = TRUE)
			location_new <- gsub(".", "", Data_Merged_by_Verify$location, fixed = TRUE)
			
			correct_st_name <- sapply(1:length(street_names),
			function(x) grepl(street_names[x], location_new[x]))
			Data_Merged_by_Verify$correct_st_name <- ifelse(correct_st_name == TRUE, 1, 0)
			
			location_words <- sapply(1:length(street_names), function(x)
			strsplit(as.character(location_new)[x], " "))
			street_type <- as.character(unique(Data_Merged_by_Verify$st_type))
			street_type <- street_type[-which(is.na(street_type))]
			
			other_st_name <- sapply(1:length(street_names), function(x)
			TRUE %in% (street_type[-(Data_Merged_by_Verify$st_type[x] == street_type)] %in%
			location_words[[x]]))
			Data_Merged_by_Verify$other_st_name <- ifelse(other_st_name == TRUE, 1, 0)

Other street name	Main data – location	1 – Yes 0 – No	If the ticket location text includes an address that is different from the location being serviced, regardless of context, returns 1. Otherwise, 0. R Code: See above (correct street names)

Table 5. Description of Engineered Features including R Code Calculations.

Engineered Feature	Source	Unit	Calculation

Phrase variables	Main data – location	1 – Yes
		0 – No
			Multiple phrase variables, each phrase with between 2 to 4 words.
			
			238 Phrases were included. They are not listed here but are listed in the Excel file which accompanies the code package entitled Phrases to Use in Modeling.xlsx.
			
			<u>R Code:</u>
			
			# Create Phrases
			phrases_used <- read_excel("Phrases to Use in Modeling.xlsx")
			phrases_used <- phrases_used$`Phrase of Phrases Violation 12.9K`
			phrase_mat <- matrix(0, nrow = nrow(Data_Merged_by_Verify),
			ncol = length(phrases_used))
			colnames(phrase_mat) <- phrases_used
			for (j in 1:nrow(Data_Merged_by_Verify)) {
			location_var <- gsub(" ", " ", Data_Merged_by_Verify$location[j])
			words <- strsplit(location_var, " ", fixed = TRUE)[[1L]]
			phrase_ngrams_2 <- vapply(ngrams(words, 2L), paste, "", collapse = " ")
			phrase_ngrams_3 <- vapply(ngrams(words, 3L), paste, "", collapse = " ")
			phrase_ngrams_4 <- vapply(ngrams(words, 4L), paste, "", collapse = " ")
			phrase_ngrams <- c(phrase_ngrams_2, phrase_ngrams_3, phrase_ngrams_4)
			phrase_mat[j,] <- ifelse(toupper(phrases_used) %in% phrase_ngrams, 1, 0)
			}

In addition, external sources of data were also considered in the feature engineering phase of modeling. These include Census Bureau population statistics, Virginia region designations, and precipitation data as shown in **Table 6**.

Table 6. List of Data Fields from External Sources

Engineered Feature	Source	Unit	Calculation

US Census population (2019)	Main data – county, Census Data File.xlsx	Integers.
		- Min: 2,190 (Highland county)
		- Max: 1,147,532 (Fairfax county)
			Obtained the US census by county data from census.gov.
			
			R code:
			
			# Census data
			Census <- as.data.frame(read_excel("Census Data File.xlsx"))
			colnames(Census) <- Census[3,]
			colnames(Census)[1] <- "Counties"
			Census <- Census[-c(1:4, 138:nrow(Census)),]
			County_names <- Census[1:133, 1]
			County_names <- sub(".", "", County_names)
			County_names <- sub(", Virginia", "", County_names)
			County_names <- toupper(sub(" County", "", County_names))
			Census[1:133, 1] <- County_names
			pop_2019 <- cbind.data.frame(county = Census$Counties,
			population = Census$`2019`)
			Data_All_New <- inner_join(Data_All_New, pop_2019, by = "county")

Precipitation	Main data
	– centroid.x, centroid.y, weather_data.xlsx
		Integers.
		• Min: 0
		• Max: 285
			Obtained the precipitation data from NOAA.
			
			R code:
			
			weather_data <- read.csv("Weather_Data.csv")
			precipitation <- rep(0, nrow(Data_Merged_by_Verify))
			weather_time <- as.POSIXlt(weather_data$time)
			for (i in 1:nrow(Data_Merged_by_Verify)) {
			comp_date <- Data_Merged_by_Verify$completed[i]
			coord <-
			c(Data_Merged_by_Verify$centroid_x[i], Data_Merged_by_Verify$centroid_y[i])
			comp_date_24hours <- c(comp_date - 3600, comp_date + 3600)
			weather_sub <- weather_data[(weather_time > comp_date_24hours[1] &
			weather_time < comp_date_24hours[2]),]
			min_station_dist <- which.min(sapply(1:nrow(weather_sub),
			function(x) geodist(coord[1], coord[2],
			weather_sub$Lon[x],
			weather_sub$Lat[x])))
			precipitation[i] <- weather_sub$precip[min_station_dist]
			}
			
			# Missing values (-9999 as indicated in data) adjusted as 0
			precipitation[which(precipitation == -9999)] <- 0
			rained <- ifelse(precipitation == 0, "No", "Yes") # Did it rain? Binary data
			
			Data_Merged_by_Verify$precipitation <- precipitation
			Data_Merged_by_Verify$rained <- rained

Rained	Main data – centroid.x, centroid.y, weather_data.xlsx	1 – Yes 0 – No	Precipitation with > 0 mm are assigned “Yes”, 0 mm assigned “No”. R Code: See above

MODELING

The goal for the statistical model is to predict the probability or percentage chance of a safety violation based on the inputs. The model was based on over 540K audited tickets from 2017-2019 and features were engineered from 1.7M tickets from 2017-2019. Once the data for modeling was designated, a variety of forms of machine learning models were considered using the R package H2O. Among these models, gradient boosted machines (GBMs) consistently outperformed multiple additional types of models. After preliminary modeling GBMs were considered exclusively for this work. A schematic of the model is shown in **Figure 5**.

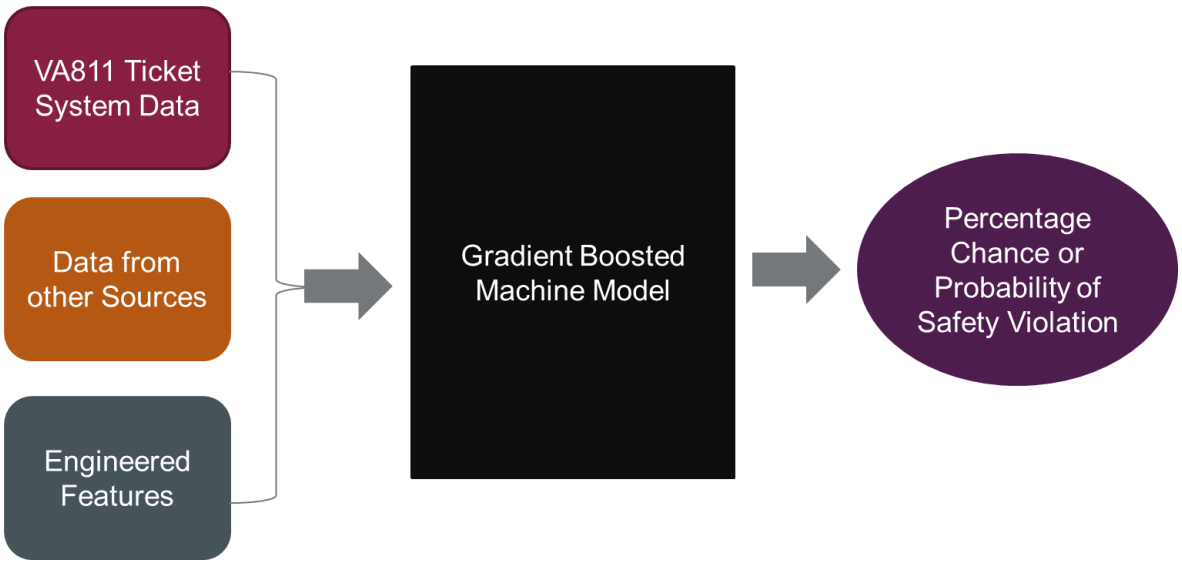


Figure 5. High Level Model Overview for Gradient Boosted Machine Model to Predict Safety Violations.

GBMs are ensemble models that fit a multiple trees iteratively on random samples of the data, selecting the next tree to compensate for the predictive weaknesses of the trees already in the model. This is accomplished by minimizing a loss function of prediction error across the trees. All trees are "averaged" to create the final prediction. A schematic of this process is shown in **Figure 6**.

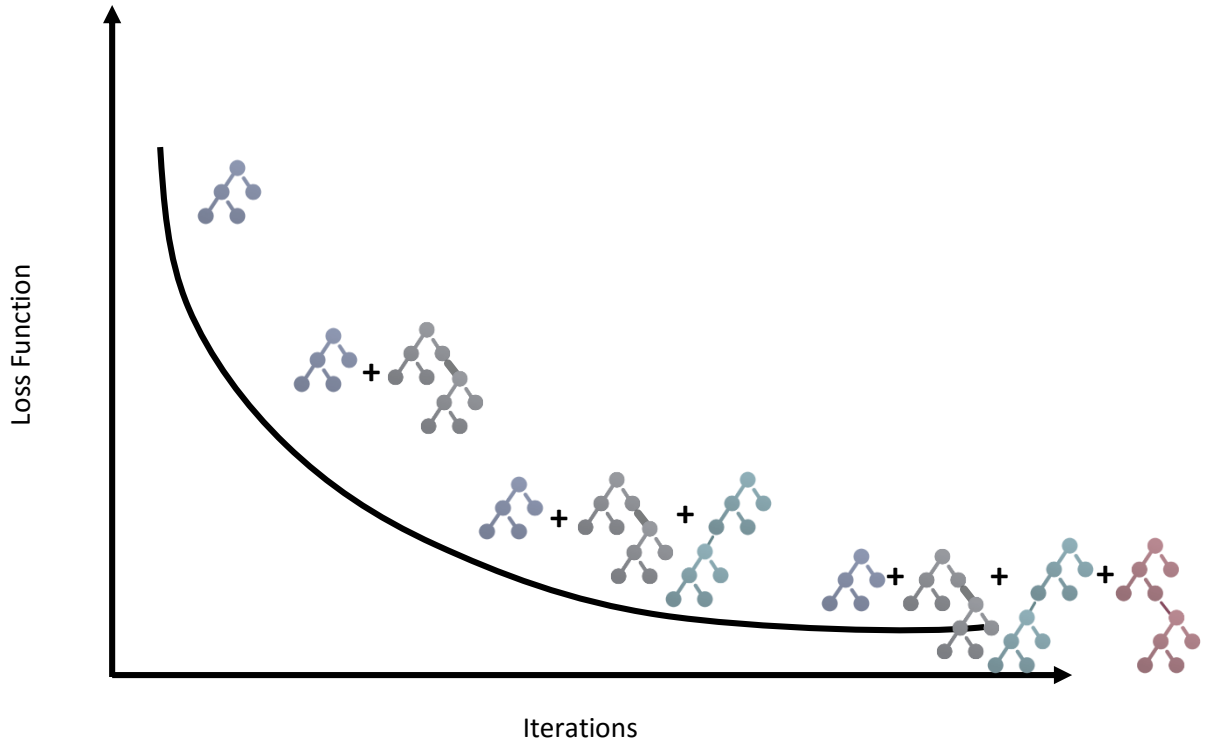


Figure 6. Schematic of Gradient Boosted Machine Model Fitting

SPECIFICS OF VA811 GBM MODEL

Models were trained on 80% of the data provided and tested on the remaining 20%. Once the final model was determined, models were fit on 100% of the data. The default parameters from the H2O R package were used.

Variable importance is used in assessing models against subject matter expertise. It provides a relative representation of the impact of each variable on predictions. **Table 7** provides a list of variables and their associated importance for variables with a relative importance > 1 from the final H2O GBM model.

Table 7. Table of Relative Variable Importances > 1 in VA811 GBM Model

Variable	Relative Importance	Scaled Importance	Percentage
ticket_age	2723.30	1.00	28.36%
county	2063.45	0.76	21.48%
violation	1226.16	0.45	12.77%
not_violation	1081.42	0.40	11.26%
number_of_tickets_by_account	644.64	0.24	6.71%

Variable	Relative Importance	Scaled Importance	Percentage
Category	345.50	0.13	3.60%
ticket_process_time	305.50	0.11	3.18%
total_ticket_num_24hrs	228.55	0.08	2.38%
time_of_day	163.61	0.06	1.70%
total_within_radius_24hrs	118.06	0.04	1.23%
ticket_completed	88.64	0.03	0.92%
age_of_account	59.93	0.02	0.62%
started	54.22	0.02	0.56%
area_in_miles	51.44	0.02	0.54%
pole ok	48.61	0.02	0.51%
match_rate	41.85	0.02	0.44%
number_characters	36.43	0.01	0.38%
ticket_operator	29.10	0.01	0.30%
number_words	26.73	0.01	0.28%
completed	26.31	0.01	0.27%
latitude	26.03	0.01	0.27%
caller_type	21.42	0.01	0.22%
longitude	19.94	0.01	0.21%
red reject	15.10	0.01	0.16%
priority	14.20	0.01	0.15%
month	13.33	0.00	0.14%
correct_st_name	11.29	0.00	0.12%
3 ft radius	9.29	0.00	0.10%
locate entire	8.01	0.00	0.08%
white flags	7.04	0.00	0.07%
emergency	6.78	0.00	0.07%
entire intersection	6.30	0.00	0.07%
left onto	5.48	0.00	0.06%
west side	4.89	0.00	0.05%
site locate	4.54	0.00	0.05%
centroid_x	4.45	0.00	0.05%
centroid_y	4.39	0.00	0.05%
revision	4.30	0.00	0.04%
north east	4.29	0.00	0.04%
right onto	4.26	0.00	0.04%
flagged route	4.05	0.00	0.04%
type	3.70	0.00	0.04%
radius around	3.69	0.00	0.04%
reject tags	2.68	0.00	0.03%
entire properties	2.63	0.00	0.03%

Variable	Relative Importance	Scaled Importance	Percentage
east side	2.58	0.00	0.03%
utility pole	2.44	0.00	0.03%
other_st_name	2.42	0.00	0.03%
power meter	2.20	0.00	0.02%
north side	2.03	0.00	0.02%
turn left	1.92	0.00	0.02%
direction_words	1.63	0.00	0.02%
red reject tags	1.19	0.00	0.01%

The important variables highlighted fall into several categories that align with business knowledge about safety violations from the subject matter experts. Account features such as the number of violations, number of tickets by account, etc. speak to the account history. Variables such as number_words, direction_words, and the phrases indicate location description complexity. Location variables such as latitude and longitude and population have long been thought to indicate likelihood of a safety violation.

USING VA811 GBM AS A CLASSIFICATION TOOL

The output of the GBM is the predicted probability of a safety violation on a scale of zero to 1 or, if you prefer to convert to percent, a percentage ranging from 0% to 100%, as illustrated in Figure 5. Yet, VA811 must establish a decision rule on this percentage to implement the model into their business process. The H2O package chooses a threshold based on maximizing critical model prediction characteristics. For the VA811 GBM model, this threshold was chosen automatically around 0.07 or 7%. However, the modeling team developed their own threshold of 0.025 or 2.5% based on principles related to the business processes and the sensitive nature of the classification outcomes. A schematic of the business classification process based on the model is shown in **Figure 7**. Further details on this decision will be provided in the Evaluation section of this report.

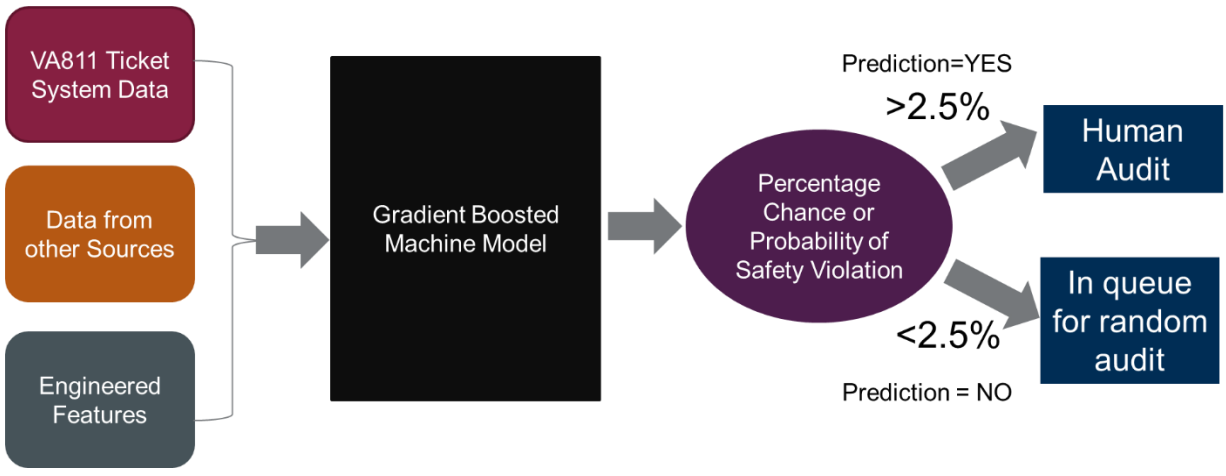


Figure 7. High Level Overview of Ticket Classification Rule from VA811 GBM Safety Violation Model

In closing the modeling section, it is important to note that gradient boosted machine models function as more of a "black box" type model. No formula can be expressed similar to those in traditional statistical modeling like regression or linear regression. No effects of specific independent variables can be described. Phenomena behind the classifications cannot be explained in terms of magnitude and direction in a general sense for these models. Evaluation of the usefulness of the models comes through variable importance and prediction quality (assessed in the evaluation section).

EVALUATION

Machine learning model quality is evaluated through a number of statistics calculated from the confusion matrix. The confusion matrix compares predicted status from the model against the actual status observed in the data. An overview of the confusion matrix is shown in **Table 8**.

Table 8. Overview of Selected Aspects of Model Performance from the Confusion Matrix

		Predicted Classification		
		NO	**YES**	**TOTAL**
Actual Classification	**NO**	True Negatives – When the actual classification is "NO", how often does the model predict "NO".	False Positives – When the actual classification is "NO", how often does the model predict "YES".	
	YES	False Negatives – When the actual classification is "YES", how often does the model predict "NO".	True Positives – When the actual classification is "YES", how often does the model predict "YES".	
	TOTAL			

Note there are many more quantities that can be calculated from the confusion matrix that indicate the predictive quality of the model. However, maximizing the true positives within the confines of reasonable audit throughput is the primary interest for this application. For that reason, additional

measures of model classification quality are not provided or discussed in this report. High performance in those other metrics would not be expected because the implementation of the model was not tailored to maximize their values.

As mentioned in the previous section on modeling, the threshold for the VA811 GBM model was selected at 0.025 or 2.5%. The resulting confusion matrix with calculation of true positives and negatives as well as false positives and negatives is shown in **Table 9**. The totals in **Table 9** are based on the total training + testing data set provided (after cleaning). By erring on the conservative side due to the safety application of the model, the model indicates the ability to identify 74% of safety violations while identifying 28% of non-safety violations as safety violations incorrectly (false positives). Models based on other thresholds that maximize overall precision or accuracy did not identify safety violations as effectively.

Table 9. VA811 GBM Calculations of Selected Aspects of Model Performance based on the Confusion Matrix

		Predicted Safety Violation		
		NO	**YES**	**TOTAL**
ACTUAL VIOLATION STATUS	**NO**	381,691 ~72% TRUE -	147,585 ~28% FALSE +	529,276
	YES	3,335 ~26% FALSE -	9,570 ~74% TRUE +	12,905
	TOTAL	385,026	157,155 29% Model Audit	542,181

DEPLOYMENT

At the writing of this report, the final model code has been handed off to VA811 and their software company Norfield for implementation. VA811 is currently piloting the model on a limited basis and Norfield is evaluating feasibility of implementation and required resources in conjunction with their other priorities.

In order to achieve maximum benefit from the model, the deployment in **Figure 8** is suggested. This deployment includes 100% audit of web entry tickets by the model as well as a random audit of negative tickets. The total audit does not exceed the current overall random audit rate of 35% of all tickets. However, one must be realistic. Changing the audit procedures to incorporate the model will also change resourcing demands and business processes. In addition, automation of this process requires a significant information technology investment. For this reason, partial model implementation may be appropriate with eventual scale up to 100% GBM model audit.

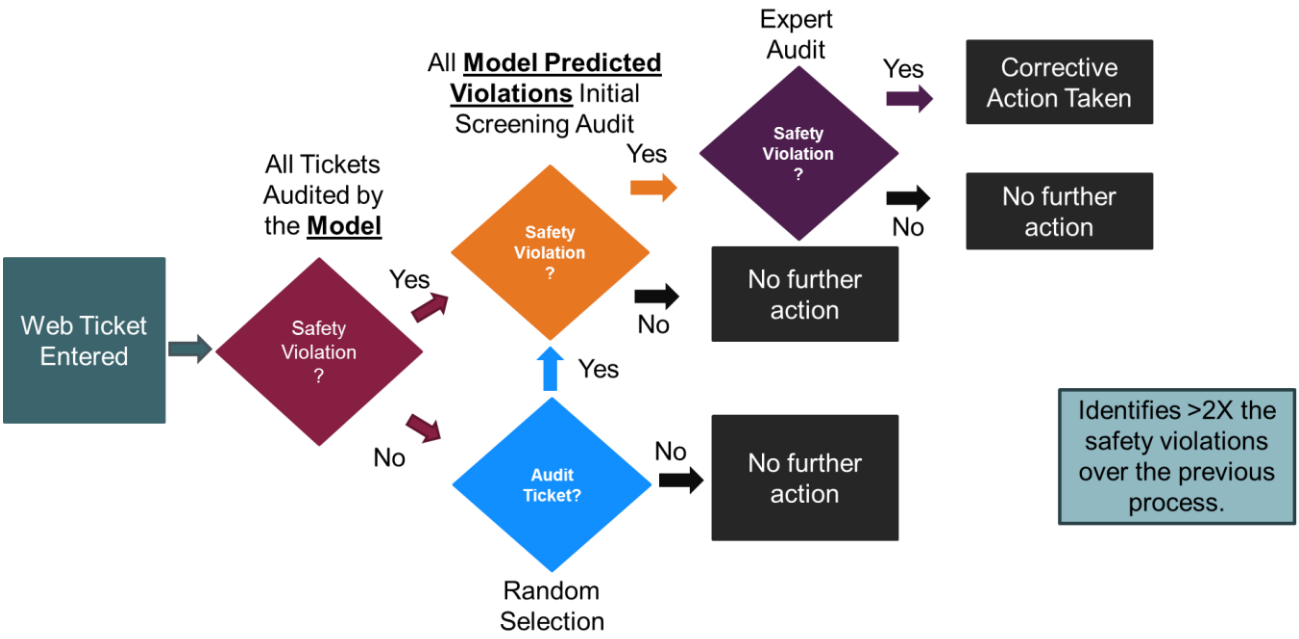


Figure 8. Deployment of VA811 GBM Model for Maximum Benefit

Figure 9 indicates an expected improvement of more than 2 times the current level of safety violations identified and corrected with 100% model audit deployment.

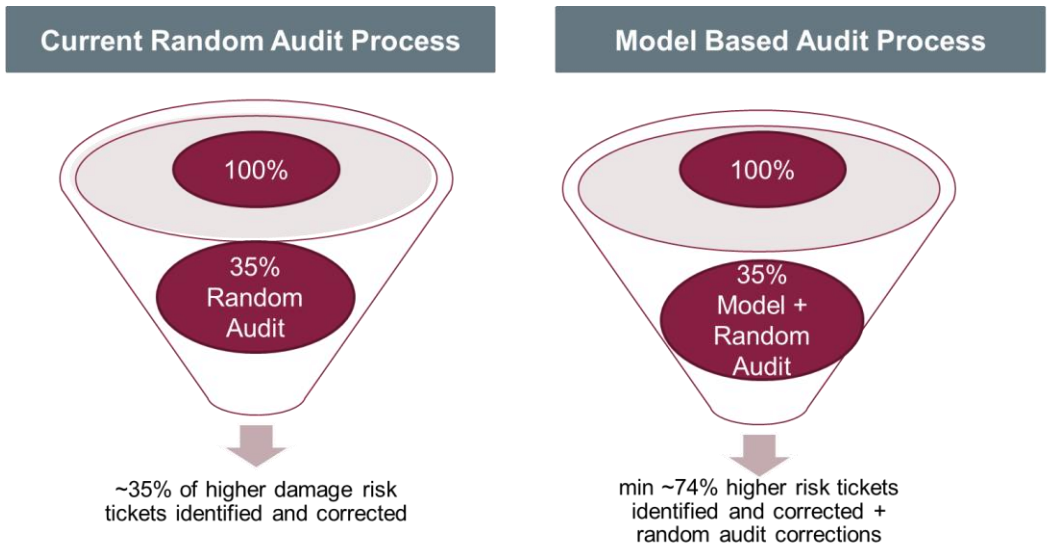


Figure 9. Expected Benefits from Model Deployment

Model deployment code has been supplied to VA811 and Norfield. This code will be submitted as part of the documentation package for the grant for review. Instructions are included to implement the model scoring code on the test data sets included in the code package.

PATH FORWARD

The analytics journey begins with dash boarding, includes predictive modeling, and ultimately artificial intelligence systems. VA811 and VT SAIG have completed important mid journey milestones in what will one day be a self-learning, self-updating artificial intelligence platform for identifying ticket damage risk for underground utilities.

The path forward begins with partial or full scale model implementation, ongoing model performance monitoring, and model updating on a regular frequency. Once proof of concept is complete, a more complex AI oriented model that captures feedback on the true state of both predicted positives and negatives is appropriate. This model will learn based on preset performance criteria and automatically update the model within the system. However, all AI systems must also be surrounded by critical thinkers that can identify when a paradigm shift is necessary.

Next generation models should incorporate additional features not currently available. These features fall into two broad categories: those requiring statistical research and those requiring enhanced data collection protocols. More research into advanced models which can determine consistency of text descriptions with map locations are required. As mentioned previously, VA811 collects data for the purpose of business processes. The data is administrative in nature. In order to become a more proactive AI organization, VA811 must strongly consider collecting data for the purposes of predictive modeling. This is a very complex task requiring definition of quantities, data governance, and investment in database and software upgrades.

CONCLUSION

As a final note, VA811 and VT SAIG have successfully completed all six phases of the defined predictive modeling project on the journey to artificial intelligence implementation at VA811. The cooperative effort hinged on shared goals, business understanding, and technical expertise. Both groups worked synergistically to provide the necessary inputs and feedback for the modeling process. The result is a model that identifies and corrects over 2 times the current safety level violations with no increased load on overall audit performance when fully deployed. Both VA811 and VT SAIG are grateful for the support of the PHMSA Grant Program that supplied the funding for this effort and look forward to future partnerships among VA811, VT SAIG, and PHMSA.

REFERENCES

Hao, K. (2018, Nov 17). *MIT Technology Review*. Retrieved 15 Sept, 2020, from technologyreview.com: https://www.technologyreview.com/2018/11/17/103781/what-is-machine-learning-we-drew-you-another-flowchart/

JMP® JMP® Version 15. (1989-2020). Cary, NC, US: SAS Institute Inc.

Lexico: Oxford English and Spanish Dictionary, Thesaurus, and Spanish to English Translator. (2020, 9 15). Retrieved from www.lexico.com: https://www.lexico.com/definition/artificial_intelligence

R Statistical Programming Language, Version 3.6.3. (2020, Sept 20). Retrieved from The Comprehensive R Archive Network: https://cran.r-project.org/

Appendix C
Quantifiable Metrics Report

AI Pilot Project Report

Kenny Spade

Date: October 5, 2020

Executive Summary

A pilot project was conducted between September 24, 2020, and October 5, 2020, to test the predictive model developed by Virginia Tech's Statistical Applications & Innovations Group (SAIG). This project involved querying VA811 database tables for current Web Ticket Entry (WTE) tickets that would be read into the program and assigned to variables in R Studio where the statistical model could then be applied. The tickets that went through this process were assigned a probability of containing a Safety Level Concern based on the model's evaluation of ticket fields and various engineered features built into the model. The first phase of the pilot project focused on auditing the tickets that exceeded VA811's designated threshold of 0.025, or 2.5%. The second phase involved auditing 20% of the tickets predicted not to contain Safety Level Concerns or tickets with probability scores below 2.5%. The results of this pilot project, as well as a number of data tables and charts, are reported below.

"Yes" Tickets

A total of 3,617 WTE tickets were read into the R program, of which 1,266, or 35%, were predicted to contain a Safety Level Concern. Those tickets were reviewed by DPS team members and the results were recorded in protected Google Sheets. The results indicated that 1,183, or 93.44%, of these tickets were either audited as Verified or a Safety Level 0 code. The remaining 6.56% of the tickets were revealed to contain a Safety Level Concern of 1 or higher. Table 1 shows the respective counts and percentages for each these Safety Level Concerns.

Safety Level	Count	%
0	1,183	93.44%
1	47	3.71%
2	4	0.32%
3	32	2.53%

Table 1. Safety Level Counts and Percentages for "Yes" Tickets

The audit codes for all Safety Level 1 or higher tickets are listed in Figure 1. Note that the audit code with the highest frequency of occurrence in this pilot was code *190 – Notification Polygon does not cover entire description of excavation–utility member(s) not missed*, which is consistently the top Safety Level Concern audit code in historical WTE data.

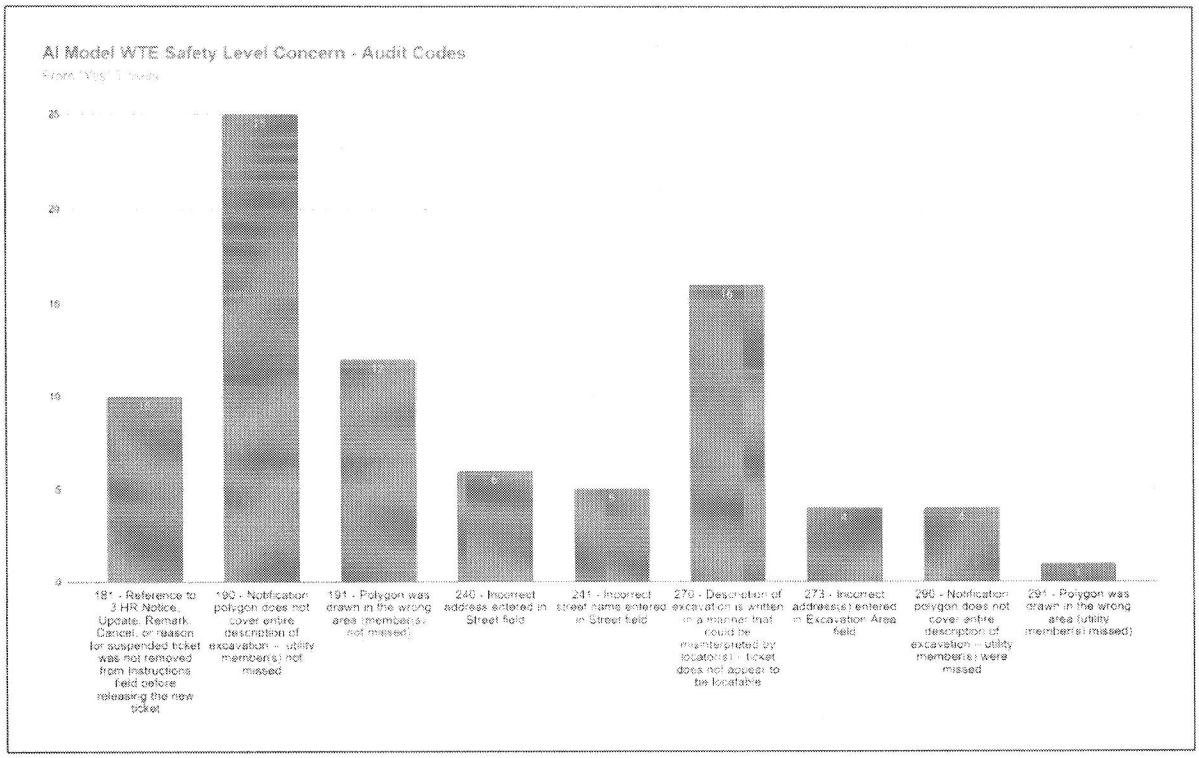


Figure 1. Safety Level 1 and higher audit codes for the "Yes" tickets.

Figure 2 shows the ranges of probability that tickets will contain a Safety Level Concern. It should be reiterated here that the threshold decided upon by VA811 was 0.025 for this model. A total of 2,351 tickets fell below this threshold.

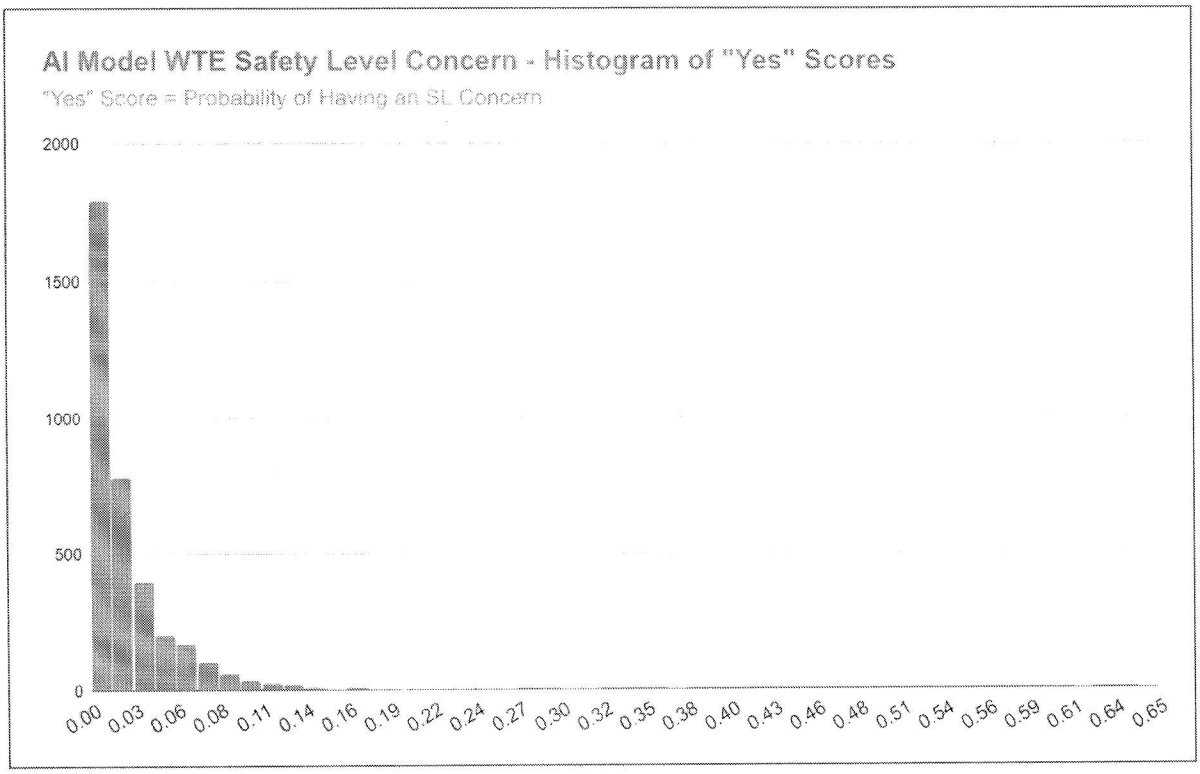


Figure 2. Histogram of probabilities that tickets will contain a Safety Level Concern.

"No" Tickets

As indicated above, a total of 2,351 tickets processed through the model were predicted not to contain a Safety Level Concern based on probability scores lower than 2.5%. VA811's plan is to audit a random sampling of these tickets. As such, the second phase of the pilot project focused on auditing a 20% sample of the tickets predicted not to contain Safety Level Concerns. A total of 470 tickets were reviewed by DPS team members during this phase. The results indicated that 453, or 96.38%, of the tickets were either audited as Verified or a Safety Level 0 code. The remaining 3.62% of the tickets were revealed to contain a Safety Level Concern of 1 or higher. Table 2 shows the respective counts and percentages for each of these Safety Level Concerns.

Safety Level	Count	%
0	453	96.38%
1	14	2.98%
2	0	0.00%
3	3	0.64%

Table 2. Safety Level Counts and Percentages for "No" Tickets.

Figure 3 shows the audit code results for the 20% of tickets predicted not to contain Safety Level Concerns. These results only include codes that are identified as Safety Level 1 or higher.

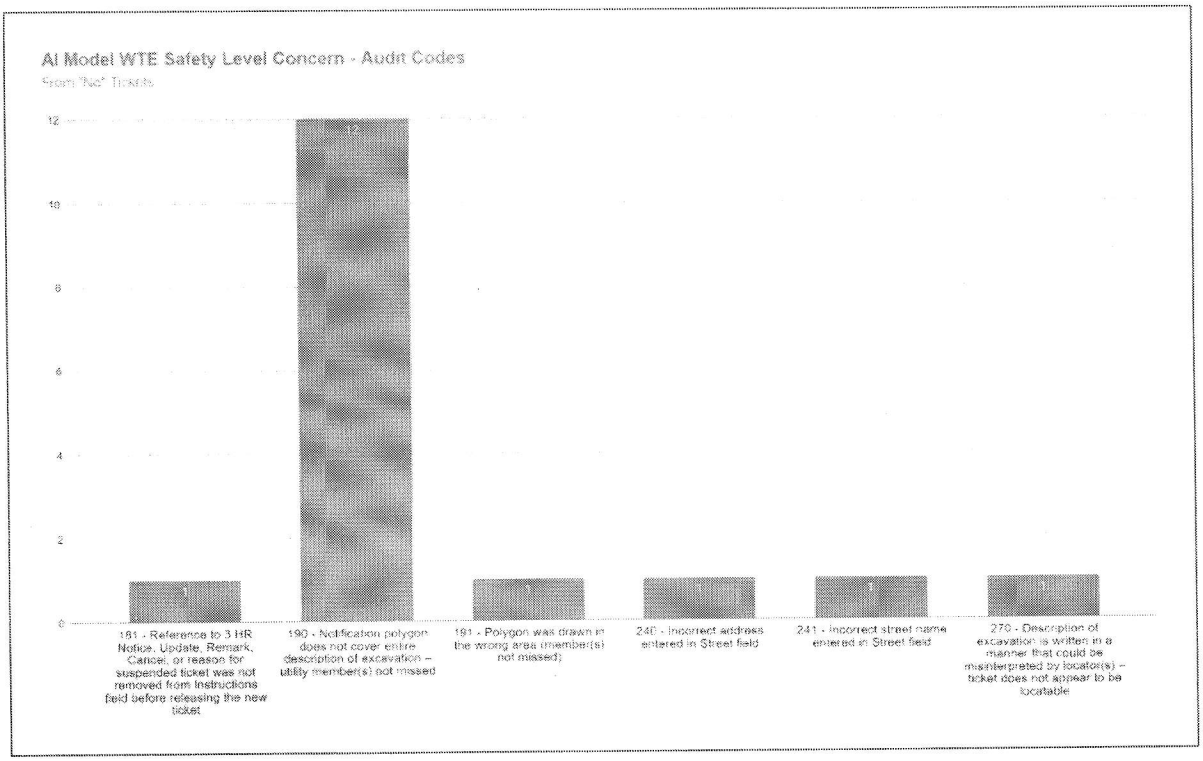


Figure 3. Safety Level 1 and higher audit codes for the "No" Tickets.

Conclusion

Throughout 2020, VA811 team members have audited between 35-45% of total WTE tickets each month (Table 3). Based on the sample of tickets processed through the predictive model in this pilot project, it is likely that any future audit program based on the model would witness a similar, if not slightly larger, number of audits being performed by team members. Historically, VA811's WTE audit program has revealed, through a combination of random audits and Risk-Based Audits, monthly Safety Level Concern rates ranging from 3.52 to 4.29% (Based on 2020 data in Table 4). The now possible 100% audits via the predictive model, combined with an estimated 35% of audits performed by team members, could nearly double the amount of errors identified each month, as demonstrated in this pilot project. This does not take into consideration the additional errors likely to be revealed through a random sampling of tickets predicted not to contain Safety Level Concerns. While many of the tickets flagged to be audited do present challenges in auditing, as many of them involve complex excavation descriptions and mapping scenarios, the benefits of using the model appear to outweigh any additional time spent auditing individual tickets. With minimal impact on daily operations, it will be possible to audit 100% of WTE

tickets, resulting in a significant increase in the number of errors that are identified and subsequently corrected by the Quality Team.

Month	Total WTE Tickets	Total WTE Audits	WTE Audit %
January	45,093	20,244	44.89%
February	45,120	18,270	40.49%
March	52,675	20,924	39.72%
April	55,286	20,845	37.70%
May	54,685	19,577	35.80%
June	57,072	19,976	35.00%
July	56,697	20,722	36.55%
August	56,709	20,490	36.13%

Table 3. Monthly WTE Audit Totals and Percentages.

Month	Total WTE Audits	Total SL	SL %
January	20,244	868	4.29%
February	18,270	669	3.66%
March	20,924	849	4.06%
April	20,845	763	3.66%
May	19,577	716	3.66%
June	19,976	800	4.00%
July	20,722	812	3.92%
August	20,490	721	3.52%
YTD Total	**161,048**	**6198**	**3.85%**

Table 4. Monthly WTE Safety Level Rates.

Appendix D Final Financial Status Report & Receipts

FEDERAL FINANCIAL REPORT

(Follow form instructions)

1. Federal Agency and Organizational Element to Which Report is Submitted	2. Federal Grant or Other Identifying Number Assigned by Federal Agency (To report multiple grants, use FFR Attachment)	Page	of
US Department of Transportation and Hazardous Materials Safety Administration	693JK31940021PSDP	1	1 pages

3. Recipient Organization (Name and complete address including Zip code)

Virginia Utility Protection Service, Inc. 1830 Blue Hills Circle, NE Roanoke, VA 24012

4a. DUNS Number	4b. EIN	5. Recipient Account Number or Identifying Number (To report multiple grants, use FFR Attachment)	6. Report Type	7. Basis of Accounting
146011619	55-0859075		☐ Quarterly ☐ Semi-Annual ☐ Annual ■ Final	☐ Cash ☐ Accrual

8. Project/Grant Period From: (Month, Day, Year)	To: (Month, Day, Year)	9. Reporting Period End Date (Month, Day, Year)
9/27/19	9/28/20	11/16/20

10. **Transactions**	Cumulative
(Use lines a-c for single or multiple grant reporting)	
Federal Cash (To report multiple grants, also use FFR Attachment):	
a. Cash Receipts	
b. Cash Disbursements	
c. Cash on Hand (line a minus b)	0.00
(Use lines d-o for single grant reporting)	
Federal Expenditures and Unobligated Balance:	
d. Total Federal funds authorized	100,000.00
e. Federal share of expenditures	99,385.87
f. Federal share of unliquidated obligations	0.00
g. Total Federal share (sum of lines e and f)	99,385.87
h. Unobligated balance of Federal funds (line d minus g)	614.13
Recipient Share:	
i. Total recipient share required	99,385.87
j. Recipient share of expenditures	
k. Remaining recipient share to be provided (line i minus j)	99,385.87
Program Income:	
l. Total Federal program income earned	
m. Program income expended in accordance with the deduction alternative	
n. Program income expended in accordance with the addition alternative	
o. Unexpended program income (line l minus line m or line n)	0.00

	a. Type	b. Rate	c. Period From	Period To	d. Base	e. Amount Charged	f. Federal Share
11. Indirect Expense							
				g. Totals:			

12. *Remarks: Attach any explanations deemed necessary or information required by Federal sponsoring agency in compliance with governing legislation:*

13. Certification: By signing this report, I certify that it is true, complete, and accurate to the best of my knowledge. I am aware that any false, fictitious, or fraudulent information may subject me to criminal, civil, or administrative penalities. (U.S. Code, Title 18, Section 1001)

a. Typed or Printed Name and Title of Authorized Certifying Official	c. Telephone (Area code, number and extension)
Scott Crawford, President & CEO	(540) 293-4292
	d. Email address: scrawford@va811.com
b. Signature of Authorized Certifying Official: B. Scott Crawford	e. Date Report Submitted (Month, Day, Year): 11/16/2020
	14. Agency use only:

Standard Form 425
OMB Approval Number: 0348-0061
Expiration Date: 10/31/2011

Paperwork Burden Statement

According to the Paperwork Reduction Act, as amended, no persons are required to respond to a collection of information unless it displays a valid OMB Control Number. The valid OMB control number for this information collection is 0348-0061. Public reporting burden for this collection of information is estimated to average 1.5 hours per response, including time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding the burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to the Office of Management and Budget, Paperwork Reduction Project (0348-0060), Washington, DC 20503.

RECEIVED
4/9/2020 hm

VIRGINIA POLYTECHNIC INSTITUTE
AND STATE UNIVERSITY
Blacksburg, VA 24061

INVOICE
1

GRANT CODE
419056

REFERENCE NO.
AT-63543

DATE: April 8, 2020

Paid 4/9/2020
Ck#17019
hm

VA811

1829 Blue Hills Circle NE
Roanoke, VA 24012

PAYMENT DUE in 30 DAYS

INVOICE MUST BE INCLUDED WITH PAYMENT
MAKE CHECKS PAYABLE TO TREASURER,VA TECH AND MAIL TO:

Office of Sponsored Programs
North End Center (MC 0170)
300 Turner Street, Suite 4200
Blacksburg, VA 24061

DATE	DESCRIPTION	CREDITS	CHARGES
10/31/19	Billing for research entitled VA811 Proposal		$10,697.76
through	VT PI - Jennifer H. Van Mullekom Sponsor Award ID: Master Research Agreement Signed 2/26/2020		
03/31/20	I certify this invoice is correct and true to the best of my knowledge.		
Partial			
	Advances	$0.00	
	Linda Goad Post Award Associate (540)231-7347		

1-30 Days Past Due	31-60 Days Past Due	Over 60 Days Past Due
$0.00	$0.00	$0.00

CURRENT AMOUNT DUE	$10,697.76
AMOUNT PAST DUE	$0.00
TOTAL AMOUNT DUE	$10,697.76

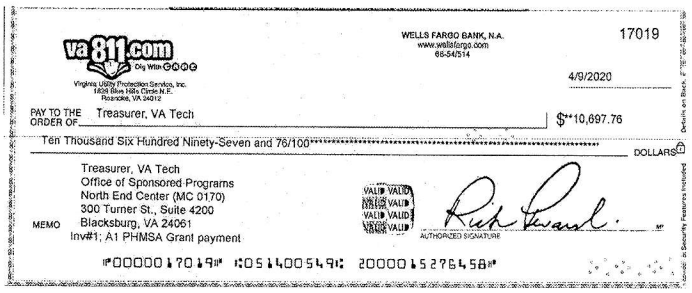
NET DUE 30 DAYS

NO CASH DISCOUNT

WELLS FARGO **Transaction Search**

Images

Date/Time Printed: 11/09/2020, 7:08 AM PST
Check 17019 - 10697.76 USD


va811.com Dig With Care
Virginia Utility Protection Service, Inc.
1828 Blue Hills Circle N.E.
Roanoke, VA 24012

WELLS FARGO BANK, N.A.
www.wellsfargo.com
65-54/514

17019

4/9/2020

PAY TO THE ORDER OF Treasurer, VA Tech $**10,697.76

Ten Thousand Six Hundred Ninety-Seven and 76/100****** DOLLARS

Treasurer, VA Tech
Office of Sponsored Programs
North End Center (MC 0170)
300 Turner St., Suite 4200
MEMO Blacksburg, VA 24061
Inv#1; A1 PHMSA Grant payment

VALID VALID VALID VALID

AUTHORIZED SIGNATURE

⑈000017019⑈ ⑆051400549⑆ 2000015276458⑈

FOR DEPOSIT ONLY
SPONSORED PROGRAMS
VPI & SU

Item Details

Account Number	**2000015276458**	Item Sequence Number	**006241808995**
Account Name	**VIRGINIA UTILITY PRO**	Bank ID	**051400549**
Check	**17019**		
Amount	**10697.76 USD Debit**		
Status	**Check Paid**		
Posting Date	**04/17/2020**		
As of Date	**04/17/2020**		

RECEIVED
5/1/2020 hm

VIRGINIA POLYTECHNIC INSTITUTE
AND STATE UNIVERSITY
Blacksburg, VA 24061

INVOICE
2

GRANT CODE
419056

REFERENCE NO.
AT-63543

DATE: May 1, 2020

Paid 5/8/2020
Ck#17072
hm

VA811

1829 Blue Hills Circle NE
Roanoke, VA 24012

PAYMENT DUE in 30 DAYS

INVOICE MUST BE INCLUDED WITH PAYMENT
MAKE CHECKS PAYABLE TO TREASURER,VA TECH AND MAIL TO:

Office of Sponsored Programs
North End Center (MC 0170)
300 Turner Street, Suite 4200
Blacksburg, VA 24061

DATE	DESCRIPTION	CREDITS	CHARGES
04/01/20	Billing for research entitled VA811 Proposal		$16,638.20
through	VT PI - Jennifer H. Van Mullekom Sponsor Award ID: Master Research Agreement Signed 2/26/2020		
04/30/20	I certify this invoice is correct and true to the best of my knowledge.		
Partial			
	Advances	$0.00	
	Linda Goad Post Award Associate (540)231-7347		

1-30 Days Past Due	31-60 Days Past Due	Over 60 Days Past Due
$0.00	$0.00	$0.00

CURRENT AMOUNT DUE	$16,638.20
AMOUNT PAST DUE	$0.00
TOTAL AMOUNT DUE	$16,638.20

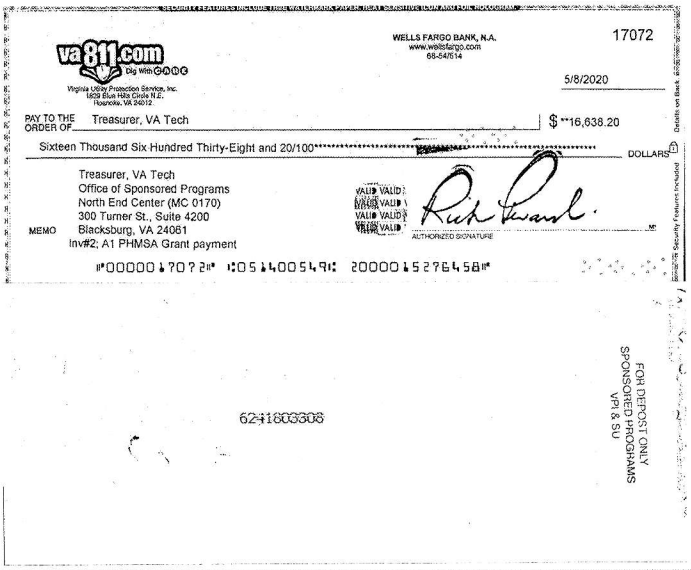
NET DUE 30 DAYS

NO CASH DISCOUNT

WELLS FARGO **Transaction Search**

Images

Date/Time Printed: 11/09/2020, 7:09 AM PST
Check 17072 - 16638.20 USD


va811.com Dig With CARE
Virginia Utility Protection Service, Inc.
1809 Blue Hills Circle N.E.
Roanoke, VA 24012

WELLS FARGO BANK, N.A.
www.wellsfargo.com
68-54/614

17072

5/8/2020

PAY TO THE ORDER OF Treasurer, VA Tech $ **16,638.20

Sixteen Thousand Six Hundred Thirty-Eight and 20/100********** DOLLARS

Treasurer, VA Tech
Office of Sponsored Programs
North End Center (MC 0170)
300 Turner St., Suite 4200
Blacksburg, VA 24061

MEMO Inv#2; A1 PHMSA Grant payment

AUTHORIZED SIGNATURE

⑈000001707?2⑈ ⑆051400549⑆ 2000015276458⑈

62418003308

FOR DEPOSIT ONLY
SPONSORED PROGRAMS
VPI & SU

Item Details

Account Number	**2000015276458**	Item Sequence Number	**006241803308**
Account Name	**VIRGINIA UTILITY PRO**	Bank ID	**051400549**
Check	**17072**		
Amount	**16638.20 USD Debit**		
Status	**Check Paid**		
Posting Date	**05/15/2020**		
As of Date	**05/15/2020**		
Additional Item Details	**CHECK** **0000005 +000000046914472**		

RECEIVED
6/1/2020 hm

VIRGINIA POLYTECHNIC INSTITUTE
AND STATE UNIVERSITY
Blacksburg, VA 24061

INVOICE
3

GRANT CODE
419056

REFERENCE NO.
AT-63543

DATE: June 1, 2020

Paid 6/5/2020
Ck#17120
hm

VA811

1829 Blue Hills Circle NE
Roanoke, VA 24012

PAYMENT DUE in 30 DAYS

INVOICE MUST BE INCLUDED WITH PAYMENT
MAKE CHECKS PAYABLE TO TREASURER,VA TECH AND MAIL TO:

Office of Sponsored Programs
North End Center (MC 0170)
300 Turner Street, Suite 4200
Blacksburg, VA 24061

DATE	DESCRIPTION	CREDITS	CHARGES
05/01/20	Billing for research entitled VA811 Proposal		$5,187.18
through	VT PI - Jennifer H. Van Mullekom Sponsor Award ID: Master Research Agreement Signed 2/26/2020		
05/31/20	I certify this invoice is correct and true to the best of my knowledge.		
Partial			
	Advances	$0.00	
	Linda Goad Post Award Associate (540)231-7347		

1-30 Days Past Due	31-60 Days Past Due	Over 60 Days Past Due
$0.00	$0.00	$0.00

CURRENT AMOUNT DUE	$5,187.18
AMOUNT PAST DUE	$0.00
TOTAL AMOUNT DUE	$5,187.18

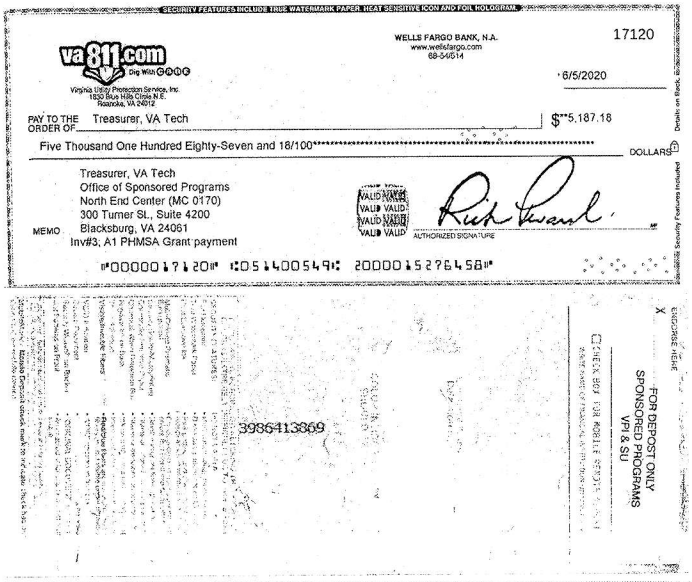
NET DUE 30 DAYS

NO CASH DISCOUNT

WELLS FARGO **Transaction Search**

Images

Date/Time Printed: 11/09/2020, 7:09 AM PST
Check 17120 - 5187.18 USD



Item Details

Account Number	**2000015276458**	Item Sequence Number	**003986413869**
Account Name	**VIRGINIA UTILITY PRO**	Bank ID	**051400549**
Check	**17120**		
Amount	**5187.18 USD Debit**		
Status	**Check Paid**		
Posting Date	**06/11/2020**		
As of Date	**06/11/2020**		
Additional Item Details	**CHECK** **0000004 +000000075444696**		

RECEIVED
7/1/2020 hm

VIRGINIA POLYTECHNIC INSTITUTE
AND STATE UNIVERSITY
Blacksburg, VA 24061

INVOICE
4

GRANT CODE
419056

REFERENCE NO.
AT-63543

DATE: July 1, 2020

Paid 7/17/2020
Ck#17221
hm

VA811

1829 Blue Hills Circle NE
Roanoke, VA 24012

PAYMENT DUE in 30 DAYS

INVOICE MUST BE INCLUDED WITH PAYMENT
MAKE CHECKS PAYABLE TO TREASURER,VA TECH AND MAIL TO:

Office of Sponsored Programs
North End Center (MC 0170)
300 Turner Street, Suite 4200
Blacksburg, VA 24061

DATE	DESCRIPTION	CREDITS	CHARGES
06/01/20	Billing for research entitled VA811 Proposal		$21,559.80
through	VT PI - Jennifer H. Van Mullekom Sponsor Award ID: Master Research Agreement Signed 2/26/2020		
06/30/20	I certify this invoice is correct and true to the best of my knowledge.		
Partial			
	Advances	$0.00	
	Linda Goad Post Award Associate (540)231-7347		

1-30 Days Past Due	31-60 Days Past Due	Over 60 Days Past Due
$0.00	$0.00	$0.00

CURRENT AMOUNT DUE	$21,559.80
AMOUNT PAST DUE	$0.00
TOTAL AMOUNT DUE	$21,559.80

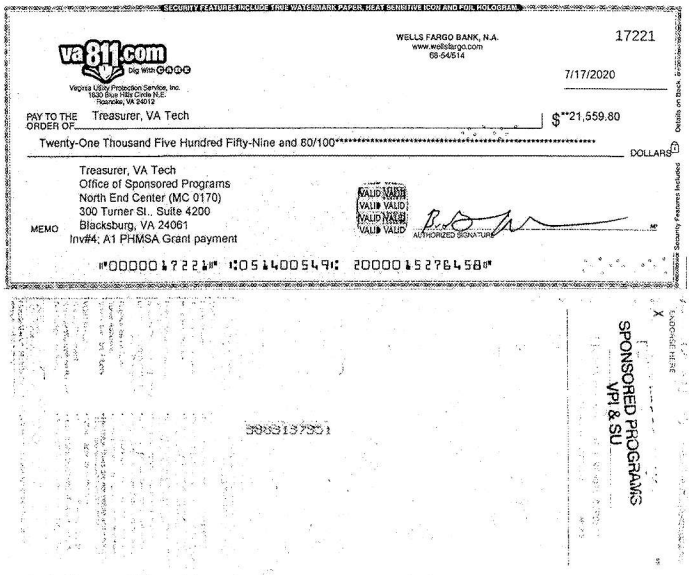
NET DUE 30 DAYS

NO CASH DISCOUNT

WELLS FARGO **Transaction Search**

Images

Date/Time Printed: 11/09/2020, 7:11 AM PST
Check 17221 - 21559.80 USD


SECURITY FEATURES INCLUDE TRUE WATERMARK PAPER, HEAT SENSITIVE ICON AND FOIL HOLOGRAM

va811.com Dig With Care
Virginia Utility Protection Service, Inc.
1630 Blue Hills Circle N.E.
Roanoke, VA 24012

WELLS FARGO BANK, N.A.
www.wellsfargo.com
68-54/514

17221

7/17/2020

PAY TO THE ORDER OF Treasurer, VA Tech $**21,559.80

Twenty-One Thousand Five Hundred Fifty-Nine and 80/100********** DOLLARS

Treasurer, VA Tech
Office of Sponsored Programs
North End Center (MC 0170)
300 Turner St., Suite 4200
Blacksburg, VA 24061

MEMO Inv#4; A1 PHMSA Grant payment

VALID VALID VALID VALID VALID VALID VALID VALID

AUTHORIZED SIGNATURE

⑈000001722⑈ ⑆051400549⑆ 2000015276458⑈

Details on Back. Security Features Included

003883137351

SPONSORED PROGRAMS
VPI & SU

Item Details

Account Number	**2000015276458**	Item Sequence Number	**003883137351**
Account Name	**VIRGINIA UTILITY PRO**	Bank ID	**051400549**
Check	**17221**		
Amount	**21559.80 USD Debit**		
Status	**Check Paid**		
Posting Date	**07/22/2020**		
As of Date	**07/22/2020**		
Additional Item Details	**CHECK** **0000001 +000000072287742**		

RECEIVED
8/3/2020 hm

VIRGINIA POLYTECHNIC INSTITUTE
AND STATE UNIVERSITY
Blacksburg, VA 24061

INVOICE
5

GRANT CODE
419056

REFERENCE NO.
AT-63543

DATE: August 2, 2020

Paid 8/6/2020
Ck#17256
hm

VA811

1829 Blue Hills Circle NE
Roanoke, VA 24012

PAYMENT DUE in 30 DAYS

INVOICE MUST BE INCLUDED WITH PAYMENT
MAKE CHECKS PAYABLE TO TREASURER,VA TECH AND MAIL TO:

Office of Sponsored Programs
North End Center (MC 0170)
300 Turner Street, Suite 4200
Blacksburg, VA 24061

DATE	DESCRIPTION	CREDITS	CHARGES
07/01/20	Billing for research entitled VA811 Proposal		$15,448.62
through	VT PI - Jennifer H. Van Mullekom Sponsor Award ID: Master Research Agreement Signed 2/26/2020		
07/31/20	I certify this invoice is correct and true to the best of my knowledge.		
Partial			
	Advances	$0.00	
	Linda Goad Post Award Associate (540)231-7347		

1-30 Days Past Due	31-60 Days Past Due	Over 60 Days Past Due
$0.00	$0.00	$0.00

CURRENT AMOUNT DUE	$15,448.62
AMOUNT PAST DUE	$0.00
TOTAL AMOUNT DUE	$15,448.62

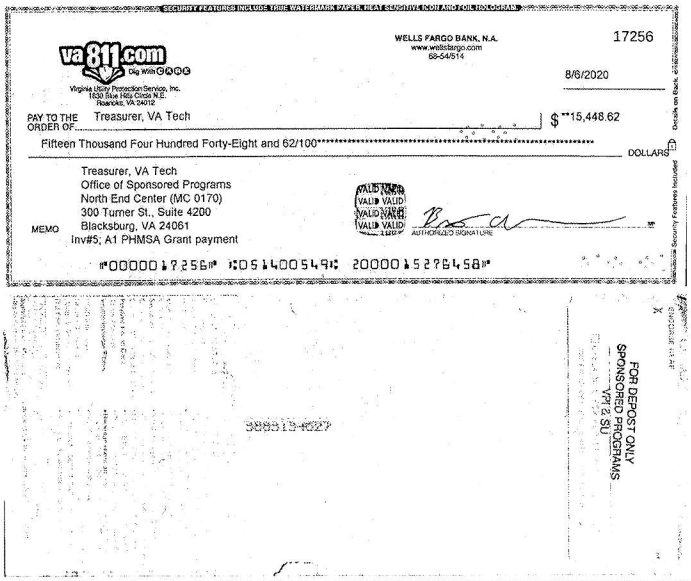
NET DUE 30 DAYS

NO CASH DISCOUNT

WELLS FARGO **Transaction Search**

Images

Date/Time Printed: 11/09/2020, 7:11 AM PST
Check 17256 - 15448.62 USD


SECURITY FEATURES INCLUDE TRUE WATERMARK PAPER, HEAT SENSITIVE INDICATOR, AND MICRO PRINTING

va811.com Dig With Care
Virginia Utility Protection Service, Inc.
1630 Blue Hills Circle N.E.
Roanoke, VA 24012

WELLS FARGO BANK, N.A.
www.wellsfargo.com
68-54/514

17256

8/6/2020

PAY TO THE ORDER OF Treasurer, VA Tech $ **15,448.62

Fifteen Thousand Four Hundred Forty-Eight and 62/100********** DOLLARS

Treasurer, VA Tech
Office of Sponsored Programs
North End Center (MC 0170)
300 Turner St., Suite 4200
Blacksburg, VA 24061

MEMO Inv#5; A1 PHMSA Grant payment

VALID VALID VALID VALID VALID VALID

AUTHORIZED SIGNATURE

⑈00001752568⑈ ⑆051400549⑆ 2000015276458⑈

Details on Back. Security Features Included

003883134627

FOR DEPOSIT ONLY
SPONSORED PROGRAMS
VA T & SU

Item Details

Account Number	**2000015276458**	Item Sequence Number	**003883134627**
Account Name	**VIRGINIA UTILITY PRO**	Bank ID	**051400549**
Check	**17256**		
Amount	**15448.62 USD Debit**		
Status	**Check Paid**		
Posting Date	**08/12/2020**		
As of Date	**08/12/2020**		
Additional Item Details	**CHECK** **0000001 +000000084685028**		

RECEIVED
9/1/2020 hm

VIRGINIA POLYTECHNIC INSTITUTE
AND STATE UNIVERSITY
Blacksburg, VA 24061

INVOICE
6

GRANT CODE
419056

REFERENCE NO.
AT-63543

DATE: September 1, 2020

Paid 9/3/2020
Ck#17302
hm

VA811

1829 Blue Hills Circle NE
Roanoke, VA 24012

PAYMENT DUE in 30 DAYS

INVOICE MUST BE INCLUDED WITH PAYMENT
MAKE CHECKS PAYABLE TO TREASURER,VA TECH AND MAIL TO:

Office of Sponsored Programs
North End Center (MC 0170)
300 Turner Street, Suite 4200
Blacksburg, VA 24061

DATE	DESCRIPTION	CREDITS	CHARGES
08/01/20	Billing for research entitled VA811 Proposal		
	VT PI - Jennifer H. Van Mullekom		$11,883.70
through	Sponsor Award ID: Master Research Agreement Signed 2/26/2020		
08/31/20	I certify this invoice is correct and true to the best of my knowledge.		
Partial			
	Advances	$0.00	
	Linda Goad Post Award Associate (540)231-7347		

1-30 Days Past Due	31-60 Days Past Due	Over 60 Days Past Due
$0.00	$0.00	$0.00

CURRENT AMOUNT DUE	$11,883.70
AMOUNT PAST DUE	$0.00
TOTAL AMOUNT DUE	$11,883.70

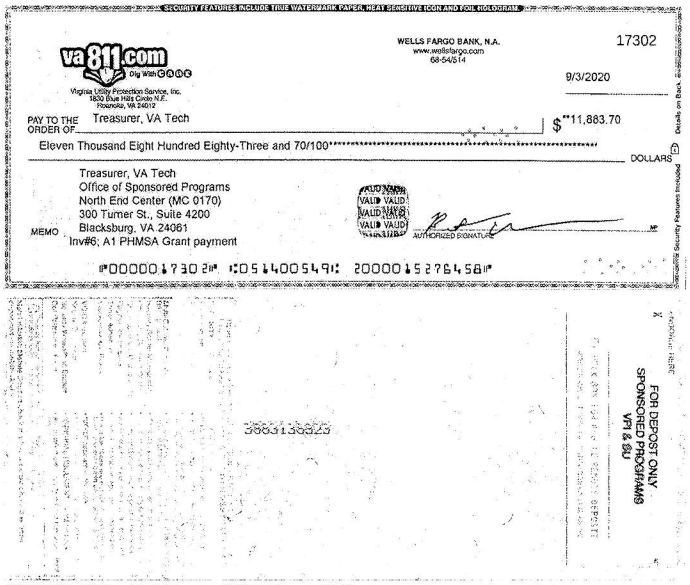
NET DUE 30 DAYS

NO CASH DISCOUNT

WELLS FARGO **Transaction Search**

Images

Date/Time Printed: 11/09/2020, 7:12 AM PST
Check 17302 - 11883.70 USD


SECURITY FEATURES INCLUDE TRUE WATERMARK PAPER, HEAT SENSITIVE ICON AND TOP, BOTTOM & EDGE

va811.com Dig With Care
Virginia Utility Protection Service, Inc.
1630 Blue Hills Circle N.E.
Roanoke, VA 24012

WELLS FARGO BANK, N.A.
www.wellsfargo.com
68-54/514

17302

9/3/2020

PAY TO THE ORDER OF Treasurer, VA Tech $**11,883.70

Eleven Thousand Eight Hundred Eighty-Three and 70/100********** DOLLARS

Treasurer, VA Tech
Office of Sponsored Programs
North End Center (MC 0170)
300 Turner St., Suite 4200
Blacksburg, VA 24061

MEMO Inv#6; A1 PHMSA Grant payment

VALID VALID VALID VALID VALID VALID

AUTHORIZED SIGNATURE

⑈000001730⑈ ⑆051400549⑆ 2000015276458⑈

Details on Back. Security Features Included

FOR DEPOSIT ONLY
SPONSORED PROGRAMS
VA TECH

Item Details

Account Number	**2000015276458**	Issue Date	**09/03/2020**
Account Name	**VIRGINIA UTILITY PRO**	Payee	**Treasurer, VA Tech**
Check	**17302**	Item Sequence Number	**003883138323**
Amount	**11883.70 USD Debit**	Bank ID	**051400549**
Status	**Check Paid**		
Posting Date	**09/15/2020**		
As of Date	**09/15/2020**		
Additional Item Details	**CHECK** **0000001 +000000118039899**		

RECEIVED
10/23/2020 hm

VIRGINIA POLYTECHNIC INSTITUTE AND STATE UNIVERSITY

Blacksburg, VA 24061

INVOICE
7

GRANT CODE
419056

REFERENCE NO.
AT-63543

DATE: October 22, 2020

Paid 10/23/2020
Ck#17390
hm

VA811
1829 Blue Hills Circle NE
Roanoke, VA 24012

Amount Paid

INVOICE MUST BE INCLUDED WITH PAYMENT
MAKE CHECKS PAYABLE TO "TREASURER, VA TECH" AND MAIL TO:

Office of Sponsored Programs
North End Center
300 Turner Street, Ste. 4200
Blacksburg, VA 24061

DATE	DESCRIPTION		
09/01/20 through 09/30/20	Billing for research entitled VA811 Proposal VT PI - Jennifer H. Van Mullekom Sponsor Award ID: Master Research Agreement Signed 2/26/2020 **I certify that all expenditures reported are for appropriate purposes and in accordance with the provisions of the application and award documents.** Linda Goad Post Award Associate (540) 231-7347		$17,970.61

1-30 Days Past Due	30-60 Days Past Due	Over 60 Days Past Due
$0.00	$0.00	$0.00

CURRENT AMOUNT DUE	**$17,970.61**
AMOUNT PAST DUE	$0.00
TOTAL AMOUNT DUE	**$17,970.61**

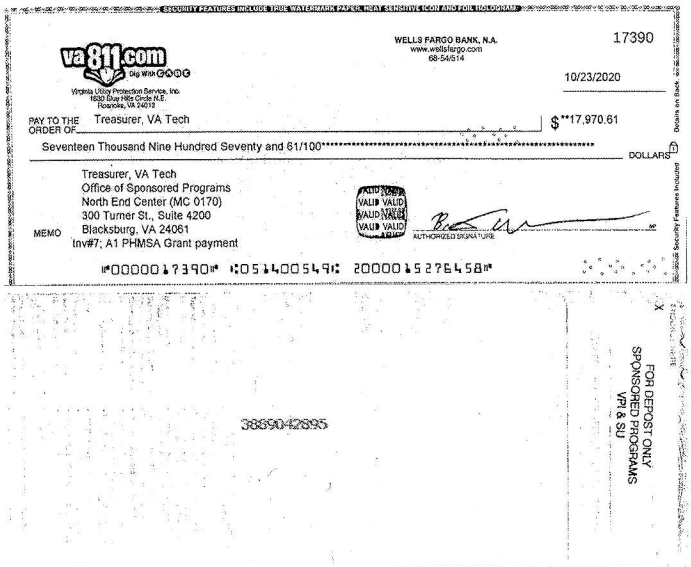
NET DUE 30 DAYS

NO CASH DISCOUNT

WELLS FARGO **Transaction Search**

Images

Date/Time Printed: 11/09/2020, 7:12 AM PST
Check 17390 - 17970.61 USD



Item Details

Account Number	**2000015276458**	Issue Date	**10/23/2020**
Account Name	**VIRGINIA UTILITY PRO**	Payee	**Treasurer, VA Tech**
Check	**17390**	Item Sequence Number	**003889042895**
Amount	**17970.61 USD Debit**	Bank ID	**051400549**
Status	**Check Paid**		
Posting Date	**10/28/2020**		
As of Date	**10/28/2020**		
Additional Item Details	**CHECK** **0000005 +000000114298426**		