Heba Maghraby, Giovanni Marrero, Joyce Yang

  Georgia Tech

## **MarketTeam Practicum: Predicting Sustainers**

## **MarketTeam Practicum: Predicting Sustainers**

Heba Maghraby, Giovanni Marrero, Joyce Yang

Applied Analytics Practicum

Georgia Tech

21 November 2024

**Introduction**

MarkeTeam is a marketing agency that collaborates with non-profit organizations to maximize donations for various important causes to improve the world. By utilizing their data analytics and marketing expertise, MarkeTeam aims to improve fundraising for these non-profit organizations by making targeted and informed decisions regarding outreach to donors.

*Problem Statement*

Our goal is to help MarkeTeam by identifying previous donors who are likely to become sustainers, or donors who are committed to donate on a regular basis, and to create a steady stream of fundraising efforts for organizations. By assessing donor data, such as transaction history, demographics, and promotional efforts by MarkeTeam, our goal is to target non-sustainers who are likely to transition to sustainers and allow MarkeTeam to specifically enhance their marketing efforts to convert these potential sustainers.

Moreover, we aim to help Markteam determine if a generic classification model for all donors is beneficial or if separate models for each donor group would be more impactful.

**Overview**

*Datasets*

MarkeTeam provides access to the following information:

1. Transaction data consisting of gifts made in the past 40 years.
2. Promotions data detailing campaign messaging received by donors.
3. Donor data detailing geographic and demographic attributes of donors.

*Clients*

We utilize data from non-profit organizations that MarkeTeam works with:

* Advocacy (League of Women Voters)
* Environmental (Sierra Club)
* Health and Health2 (Children's Hospital LA and Vanderbilt U Medical Center)
* Social (UNICEF)
* Veteran (Paralyzed Veterans of America)

The counts of gifts, donors, and promotions by client are provided in the table below.

|  |  |  |  |
| --- | --- | --- | --- |
| Client | Gifts | Donors | Promotions |
| Environmental | 14,608,861 | 1,987,044 | 150,998,930 |
| Social | 11,304,354 | 1,842,730 | 101,119,876 |
| Veteran | 50,576,096 | 3,553,488 | 296,795,118 |
| Advocacy | 1,062,540 | 275,701 | 17,048,227 |
| Health | 907,348 | 138,062 | 231,850 |
| Health2 | 227,131 | 35,306 | 6,454,338 |

Our goal in this project is to develop a technique to predict sustainer donor conversion based on information about the donors and promotions they receive, both for Generic (not client-specific) gifts and gifts specific to the Veteran client.

We do this by identifying the date of conversion and the donor’s behavior 24 months prior to conversion.

*Data Wrangling*

A Snowflake warehouse was established to manage a substantial amount of data (approximately 45 gigabytes) using SQL, enabling the preparation of a clean dataset suitable for machine learning applications. Initially, all non-USA donors were excluded from the tables, after which all six data tables were combined to create a unified dataset. The Veteran dataset was retained to be utilized in a specific Veteran sub-model.

Next, the sustainers' pool was identified using the PLEDGE\_SUSTAINER variable, and a new variable, SUS\_FLAG, was created based on this identification. Subsequently, one-time donors were recognized, leading to the creation of the ONE\_TIME\_FLAG variable. The conversion date was then established, and the history of sustainers was narrowed down to the 24 months before their conversion. Similarly, data for non-sustainers was limited to the 24 months preceding their last donation date. Finally, outliers were identified as values exceeding three standard deviations from the mean, creating an OUTLIER\_FLAG variable.

After reducing the data set, the following features were engineered:

|  |  |
| --- | --- |
| **Feature** | **Description** |
| DISTINCT\_APPEAL\_COUNT | the total number of promotion codes sent to each donor |
| RESPONSE\_RATE | The response rate to appeal codes sent in promotions |
| AVG\_DAYS\_ACROSS\_APPEAL\_CODES | The average days between the date a promotion code was sent and the donation date |
| MOST\_COMMON\_GIFT\_CHANNEL\_PRO | The most common gift channel via promotion |
| DONATION COUNT | The number of times a donor donated |
| MAX\_GIFT | Maximum amount gift per donor |
| AVG\_GIFT | The average of donations per donor |
| TOTAL\_GIFT | Total amount gifted per donor |
| AMOUNT\_GIFTED\_LAST\_YEAR | The amount gifted in the prior 12 months per donor |
| TOP\_5\_SUS\_PERCENT\_FLAG | A flag determining if the donor resides in one of the top 5 states where sustainers live |

A full list of features used in the models is available in Appendix 1.

*Handling Class Imbalance*

One of the biggest challenges in this dataset is the significant imbalance, as less than 1% of the donor population is classified as sustainers. A few approaches were implemented to address this imbalance:

* Undersampling: We reduced the non-sustainers pool to a ratio of 70%: 30% for sustainers by omitting any rows with missing values in AGE, DISTINCT\_APPEAL\_COUNT, AVG\_DAYS\_ACROSS\_APPEALCODES, RESPONSE\_RATE, TOTAL\_AMOUNT\_GIFTED\_PRO, or AVG\_DAYS\_ACROSS\_APPEALCODES. Then, we performed random sampling from the resulting dataset. The remaining data was used.
* Oversampling: In the Veterans dataset, which is smaller, the oversampling approach was used to balance the sustainers pool to a 50%:50% ratio.

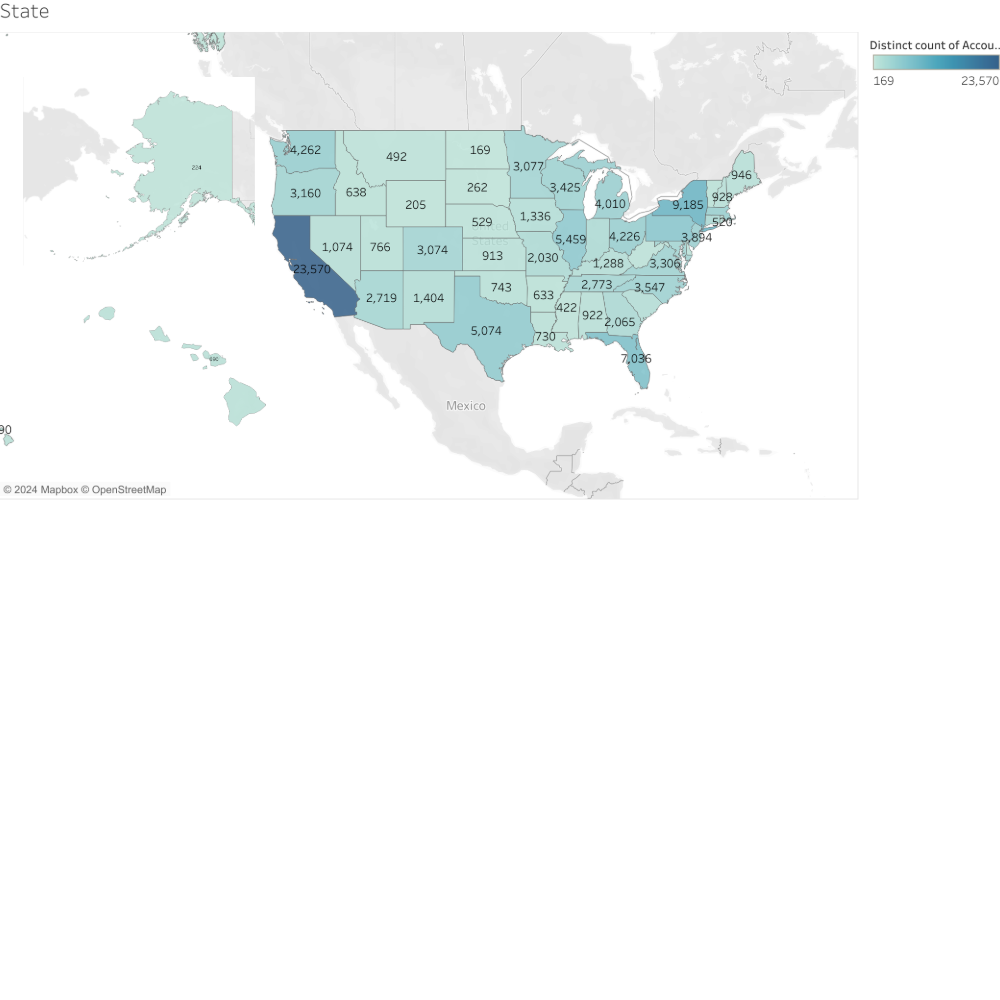
*Handling Missing Values*

Rows with missing values in the non-sustainer pool were removed. However, due to the small size of the sustainer pool, we applied data imputation. Missing age values were replaced with the average age, while intuitive values or "unknown" were used for the other columns.

**Exploratory Data Analysis (EDA) Summary: Generic Model Data**

We explored the distribution of variables in the generic (not client-specific) data.

1. **Geographic Distribution**

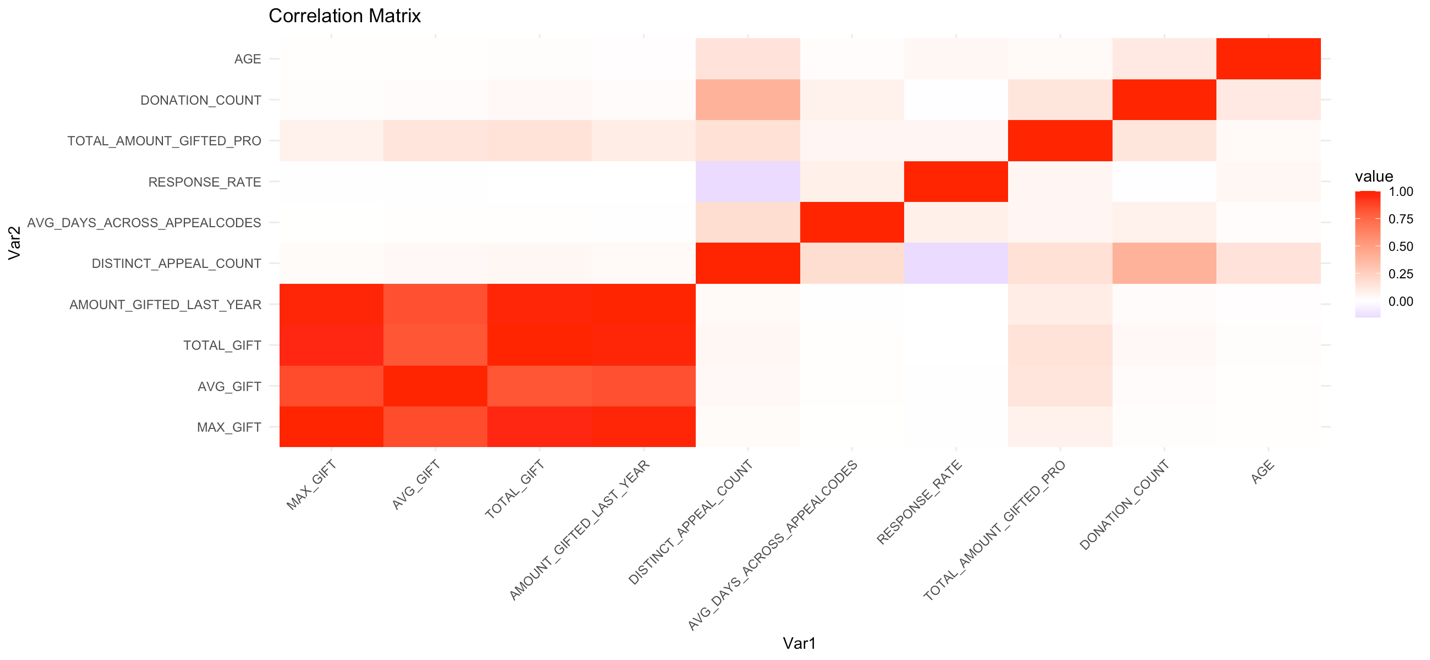


1. **Numeric Variables**

The exploratory analysis of the dataset indicates that most of the distributions are highly right skewed (graphs available in appendix 2), which suggests that most donors contribute modest amounts or interact infrequently, while a small subset contributes significantly more or interacts more often. Below is a summary of key findings across various metrics:

* **AGE**: The dataset is composed primarily of older individuals, with most AGE values between 70 and 80 years. There are fewer younger individuals, and a small number of individuals over 100, which may represent outliers.
* **DONATION\_COUNT**: Most donors make 0-10 donations, indicating infrequent contributions, while very few make over 100 donations. A skewed distribution shows that a small group of donors contributes at a much higher frequency.
* **TOTAL\_AMOUNT\_GIFTED\_PRO**: Contributions are highly skewed, with most donors contributing minimal amounts and a few making large donations up to 100,000. Focusing on smaller values, most contributions are below 100.
* **RESPONSE\_RATE**: Most accounts have low response rates, with values concentrated near zero. High response rates are rare, leading to an extreme right-skew.
* **AVG\_DAYS\_ACROSS\_APPEALCODES**: Most values are near zero, indicating frequent interactions within a short time span, though some records show extremely long intervals.
* **DISTINCT\_APPEAL\_COUNT**: Most donors receive a small number of unique appeals, with a smaller subset receiving more varied campaigns, indicating targeted outreach.
* **AMOUNT\_GIFTED\_LAST\_YEAR:** Contributions are mostly low, with a few large gifts extending up to 1,000,000. Most donations are clustered below 200.
* **TOTAL\_GIFT:** Most donations are concentrated at the lower end of the spectrum, with very few exceeding 100.
* **AVG\_GIFT**: Typical average donations per donor are low, with most values below 50.
* **MAX\_GIFT:** Most maximum gifts are small, with only a few exceptionally large donations up to around 1,000,000. Most maximum gifts are below 100.

The data consistently shows right-skewed distributions across various metrics, indicating that most donors contribute smaller amounts or have limited interactions, while a few donors give substantial sums or engage more frequently. Addressing this skewness through data transformations or segmentation is necessary to avoid biases in subsequent analyses or models. To transform the data, we take the log of the TOTAL\_AMOUNT\_GIFTED\_PRO variable.



Strong positive correlations are observed among TOTAL\_GIFT, AMOUNT\_GIFTED\_LAST\_YEAR, AVG\_GIFT, and MAX\_GIFT. Additionally, DONATION\_COUNT shows notable positive correlations with these variables, indicating that individuals who donate more frequently tend to contribute larger total amounts. Moderate correlations are evident between DISTINCT\_APPEAL\_COUNT and TOTAL\_AMOUNT\_GIFTED\_PRO.

Interestingly, a negative correlation between RESPONSE\_RATE and DISTINCT\_APPEAL\_COUNT suggests that donors are less likely to respond as the variety of appeals they receive increases. In contrast, AGE exhibits weak or no correlation with most variables, indicating that donor age may not significantly impact donation patterns.

1. **Categorical Variables**

Most donors are not couples; in addition, female donors are predominant. The ONE\_TIME\_FLAG and OUTLIER\_FLAG variables show a clear majority of 0, indicating that most donors are classified neither as one-time donors nor as outliers. For the TOP\_5\_SUS\_PERCENT\_FLAG, most of the values are also 0, suggesting that relatively few donors are from the top five states by percentage of donors (graphs available in appendix 1).

**Methods**

To ensure that we do not retroactively assign sustainership based on sustainer donations, we limit our datasets to only activity before the 'PledgeSustainer' flag became true.

We use SQL to separate the sustainers from non-sustainers. The count for sustainers and non-sustainers is shown below.

|  |  |  |
| --- | --- | --- |
| CLIENT | SUSTAINERS | NON-SUSTAINERS |
| ADVCY | 3,377 | 272,324 |
| ENVR | 7,915 | 1,979,129 |
| HEALTH | 4,052 | 134,010 |
| HEALTH2 | 339 | 34,967 |
| SOCIAL | 84,733 | 1,757,997 |
| VETERAN | 33,516 | 3,519,972 |

*Datasets used*

1. Generic model dataset (70% non-sustainers: 30% sustainers).
2. Veteran model dataset (70% non-sustainers: 30% sustainers)
3. Oversampled Veteran dataset (50% non-sustainers: 50% sustainers).
4. Veteran dataset with no imbalance handling (99% non-sustainers: 1% sustainers)

*Models*

Given the predictive nature of the problem, as well as the binary outcome (sustainer or non-sustainer), we use the six modeling approaches below:

|  |  |  |
| --- | --- | --- |
| Logistic Regression | Random Forest | Boost |
| Support Vector Machine | Naive Bayes | Neural Network |

**Logistic Regression**

We ran the logistic regression with and without outliers and found that the model without outliers performed better for the Generic case by both recall and accuracy. For the sake of time, we omitted outliers for the Veteran case.

For variable selection, we used stepwise regression and determined the optimal variable subset using AIC. We found that this dropped some variables such as MAX\_GIFT and MOST\_COMMON\_GIFT\_CHANNEL indicator variables. For both Generic and Veteran cases, the logistic regression did not perform among the top 50% of models by accuracy and recall. The poor performance may be due to the performance-interpretability tradeoff; while logistic regression is interpretable compared to more opaque models such as neural networks, it lacks the flexibility of the more complex models, especially in the scenario with outliers.

**Support Vector Machine**

Support vector machine (SVM) is a classifier that partitions data points based on a hyperplane. Data on one side of the hyperplane are predicted to belong to a class, and data on the other side are predicted to belong to the other class.

The support vector machine model was run with and without outliers; the model without outliers performed significantly better. As with logistic regression, we excluded outliers for the Veteran model. To select variables, we used the same subset of variables that was determined for logistic regression. For the Generic case, the support vector machine performed well compared to the other methods, achieving a recall (true positive rate) of 1; however, for the Veteran case, SVM performed in the lower half of models.

**Random Forests**

We used random forest as a modeling approach for both of our datasets. Random forest is a very robust machine learning algorithm that is used for classification that is built on several decision trees by generating several trees and then aggregating their outputs to improve model performance and reduce potential overfitting. We expect that random forest would be a successful modeling approach to our problem, since it is very powerful for working with complex datasets that possess many predictions.

In our Generic models, we developed models that were trained on two datasets, one with outliers and one that removed outliers. We also tested both models on testing data that was with and without outliers as well. We observed that the model that was trained on data with outliers led to much better performance when predicting on both test datasets. We anticipate that our outliers may be beneficial in providing modeling efforts with helpful datapoints.

In our Veteran models, we see that random forest modeling approached were the best performing model. We attempted modeling in several different ways in our approach. We generated random forest modeling with as much data as our systems could handle, which was ultimately only 20% of the train data for our Vet2 dataset. Even with limited data to train, this model performed with a perfect recall, offering us a testing error of 0.006. We also created models based with variable selection based on Cumulative Importance, with thresholds of 70%, 80%, & 90%. We observed that the performances of models improved as we increased the threshold for which variables were included, which leads us to believe that utilized all variables was the most beneficial approach in random forest modeling. Ideally, with more computational capacity, we would be able to generate a random forest model that utilizes all the training data, which we suspect would provide the best possible random forest model for our approach on Veteran datasets, and possibly the best overall model for this dataset.

**Naïve Bayes**

We also used Naïve Bayes classifiers as a modeling approach for both of our datasets. Naïve Bayes typically performs best on data that is given categorical data. This machine learning algorithm assumes that the features in the given dataset are independent of each other, which we do not see in either our Generic or Veteran datasets. We did not anticipate this model to perform very well given its limitations to handling quantitative data.

For both the Generic & Veteran models, we attempted to apply Laplace smoothing as well to alter parameters for better model performance, but we ultimately noticed that these models performed very poorly, both generally & in relation to our other models. We noted that they can offer high values of true positive results, but that is because they typically over-classified Sustainers in Generic while returning very high training and testing errors.

**Gradient Boosting (LightGBM)**

The gradient boosting algorithm is robust in handling mixed data types and achieving high accuracy with little tuning. Additionally, it can handle categorical variables directly. Moreover, it is excellent for large datasets and handles non-linear relationships well.

The algorithm was applied to the generic dataset with outliers and without outliers. It was also tested using all variables and on a selection of variables based on their relative importance (a complete description of the models used, and their performance matrix can be found in Appendix 3). A collection of parameters was tested, including 100, 500, 1000, 5000 & 10,000 trees. The model trained on the dataset with no outliers, then tested on the dataset with outliers and had 10,000 boosting iterations, was the best-performing model with a 99.13% Recall and 99.74% Accuracy.

The same technique was applied to the Veteran dataset in addition to testing the datasets with a 50:50 class balance and 99:1 class balance. The best-performing model was trained and tested on the 70:30 dataset and 500 boosting iterations, with a Recall of 99.36% and an Accuracy of 96.04%.

Overall, the datasets with 50:50 and 99:1 class ratio did not perform well under this algorithm. Moreover, the generic model outperformed the sub-model in this algorithm.

**Neural Networks**

Neural networks can handle complex relationships between features and the target, making them flexible. Moreover, they can capture complex patterns and adapt to various data types.

The algorithm was applied to the generic dataset with and without outliers and tested using all variables. A collection of parameters, including 2,5 & 10 hidden layers and 100, 200, and 500 iterations, was tested. The model trained on the dataset with no outliers was then tested on the dataset with outliers, had 10 hidden layers and 500 iterations, and was the best-performing model, with a 99.06% Recall and 98.12% Accuracy.

The same technique was applied to the Veteran dataset in addition to testing the datasets with a 50:50 class balance and 99:1 class balance. The best-performing model was trained and tested on the 70:30 dataset with no outliers, then tested on the same dataset with outliers with a Recall of 96.46% and Accuracy of 96.7%. The parameters used for this model were 10 hidden layers and 500 iterations.

Overall, the 50:50 & 99:1 class ratio dataset didn’t perform well under this algorithm. Moreover, the generic model outperformed the sub-model in this algorithm. Additionally, the 99:1 model had a very low F-1 score of 37.88%.

**Model Performance**

We chose to observe Recall & Accuracy as key metrics in determining model performance. Specifically, we utilized Recall based on the importance of returning True Positive values with a low cost of also including False Positives. It would be very costly for MarkeTeam to exclude potential Sustainers all together when targeting them for transition, so we have a strong preference for having False Positives as opposed to False Negatives. We also used Accuracy to provide a general sense of correctness, to make sure that the model is performing well overall and not providing a very low threshold for determining potential Sustainers.

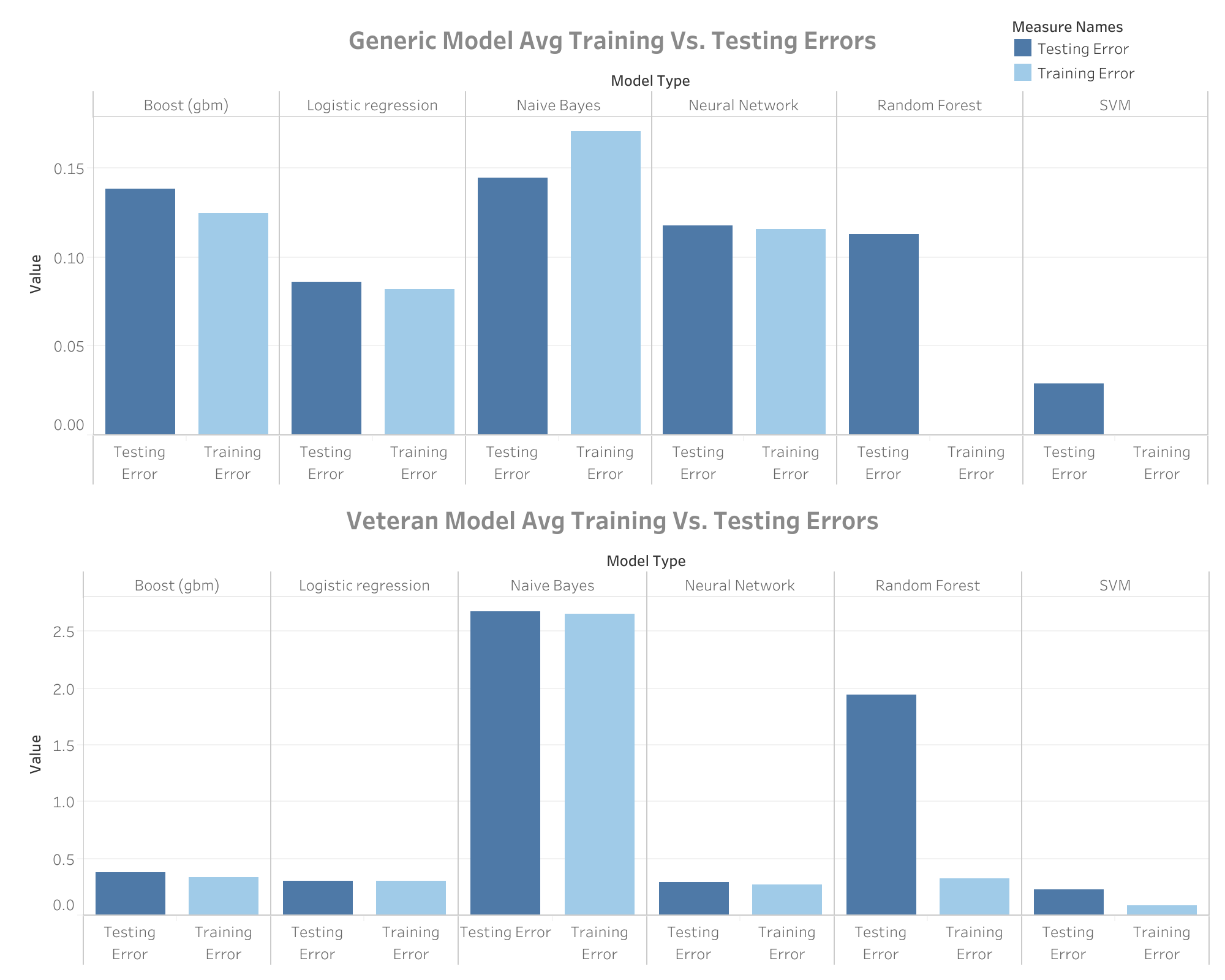
**Detailed Performance Metrics (Generic Model)**

|  |  |  |
| --- | --- | --- |
| **Model with Best Recall** | **Best Recall** | **Accuracy** |
| Support Vector Machine | **0.9989** | **0.9997** |
| Boost | 0.9969 | 0.9817 |
| Neural Network | 0.9943 | 0.9678 |
| Naive Bayes | 0.9664 | 0.9280 |
| Random Forest | 0.9424 | 0.9823 |
| Logistic Regression | 0.9146 | 0.9718 |

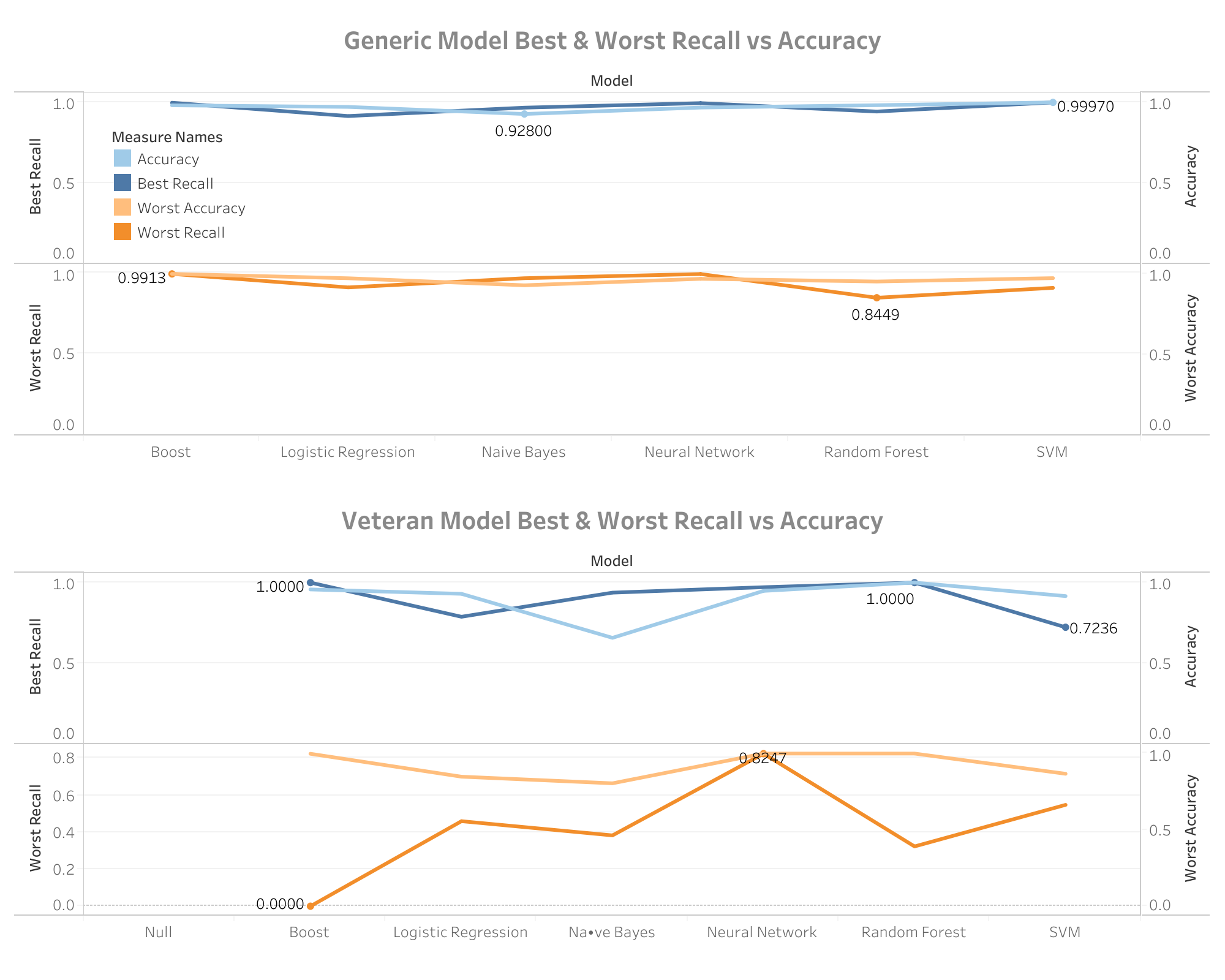
**Detailed Performance Metrics (Veteran Model)**

|  |  |  |
| --- | --- | --- |
| **Model with Best Recall** | **Best Recall** | **Accuracy** |
| Random Forest | **1** | **0.9994** |
| Boost | **1** | 0.9571 |
| Neural Network | 0.9715 | 0.9481 |
| Naive Bayes | 0.9379 | 0.6578 |
| Logistic Regression | 0.7886 | 0.9300 |
| Support Vector Machine | 0.7236 | 0.9160 |

**Models Comparison**



The generic model had significantly lower training and testing errors. The Naive Bayes model was the worst-performing model in both datasets, while SVM was the best-performer.

In the generic model, accuracy across models is consistently high, with the best accuracy near 1.0 for all models and the worst recall slightly varying but remaining above 0.8. The Boost model shows a slight dip in worst recall (0.8449). In the veteran model, performance is more varied, with models like Neural Network and Logistic Regression achieving perfect accuracy (1.0) and high recall, while others like Boost and SVM display lower worst recall values. This comparison highlights the stability and reliability of models like Neural Networks and Logistic Regression in both datasets while exposing potential limitations in Boost and SVM models, particularly in the veteran dataset.

**Recommendations**

As shown in the previous section, different models work well for different cases. For the Generic Model, Support Vector Machine and Boost have the highest recall (true positive rate) while for the Veteran Model, Random Forest and Boost have the best performance, with lower performance for the interpretable models (logistic regression and SVM).

**Challenges & Limitations**

Throughout this project, one of the most common challenges encountered was with handling very large datasets. Due to the size of the datasets, we had several difficulties in manipulating data and generating models. We found several ways to incorporate cloud computing software, such as Snowflake, to manipulate and trim down large datasets into a size that was workable locally in R.

When generating models, specifically in the more computational expensive modeling techniques, such as random forest, we had to find ways that utilized as much data as possible in both determining training datasets and generating the models themselves. We operated on a typical 70-30 split for training and testing data but sometimes would not be able to run computationally expensive models on these large datasets, so we would trim this using random splits again. We also found that in random forest modeling, we were not able to use all variables, so instead we utilized cumulative feature importance in variable selection, trying several different thresholds to use as much data as possible in addition to the cumulative feature threshold.

**Conclusion**

The scope of the project encompassed developing General and Veteran models to identify sustainers. Ultimately, we determined that when creating predictive models for identifying sustainers in the Generic dataset, we had our best performances in SVM, Boost, and Neural Network models. In our Veteran dataset, we observed that our Random Forest, Boost, and then Neural Network models were the best performers.

As expected with this problem, black box algorithms were typically the best performing models due to the original datasets having both categorical and numerical data, which was often highly correlated. Overall, we noticed that we were able to get better models in our Generic dataset but ultimately had the best performing model in our Veteran dataset. We would highly suggest that MarkeTeam should increase computational capacity to afford the ability to run additional black box algorithms such as Neural Networks, Boost, and Random Forest to get the best performance out of their modeling efforts to predict potential sustainers. We suspect that specialized models for specific organizations would be the best way to find potential sustainers in the future, while there is certainly value in still using the Generic approach as well.

**Acknowledgments**

We would like to thank our collaborators at MarkeTeam for the opportunity to work with them on this project and offering guidance for any questions regarding the datasets and business intentions. We would also like to thank Georgia Tech for the opportunity to partner us with MarkeTeam and providing the background knowledge to provide helpful insights with real-world applications.

**Summary of Work**

|  |  |  |  |
| --- | --- | --- | --- |
| Task | Description | Team Member Contributions | Target Date |
| Data Cleaning | Identify sustainers and split data | HM,GM,JY | 9/27/24 |
| Geographical Analysis | Create top 5 flags | JY | 9/28/24 |
| Exploratory Data Analysis | Examine variables for model | GM | 9/28/24 |
| Outlier Analysis | Detect outliers using Cook's distance | HM | 10/1/24 |
| Generic Model Building | Use regression, SVM, RF, NB, Boost, neural networks | HM,GM,JY | 10/14/24 |
| Veteran Model Building | Use regression, SVM, RF, NB, Boost, neural networks | HM,GM,JY | 11/08/24 |
| Detailed Performance Metrics | Describe how variables and modeling differ by model | HM,GM,JY | 11/15/24 |
| Final Report | Consolidate analysis and write final summary | HM, GM, JY | 11/21/24 |

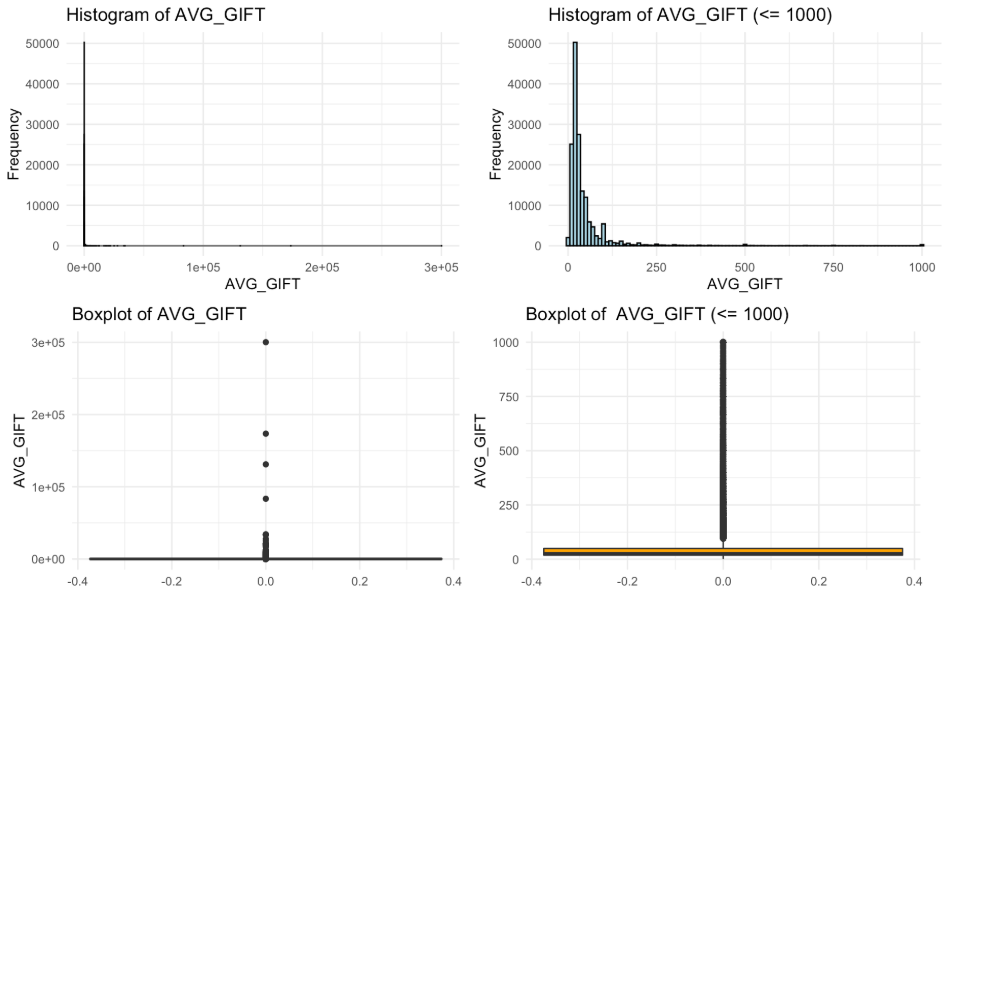
**Reference**

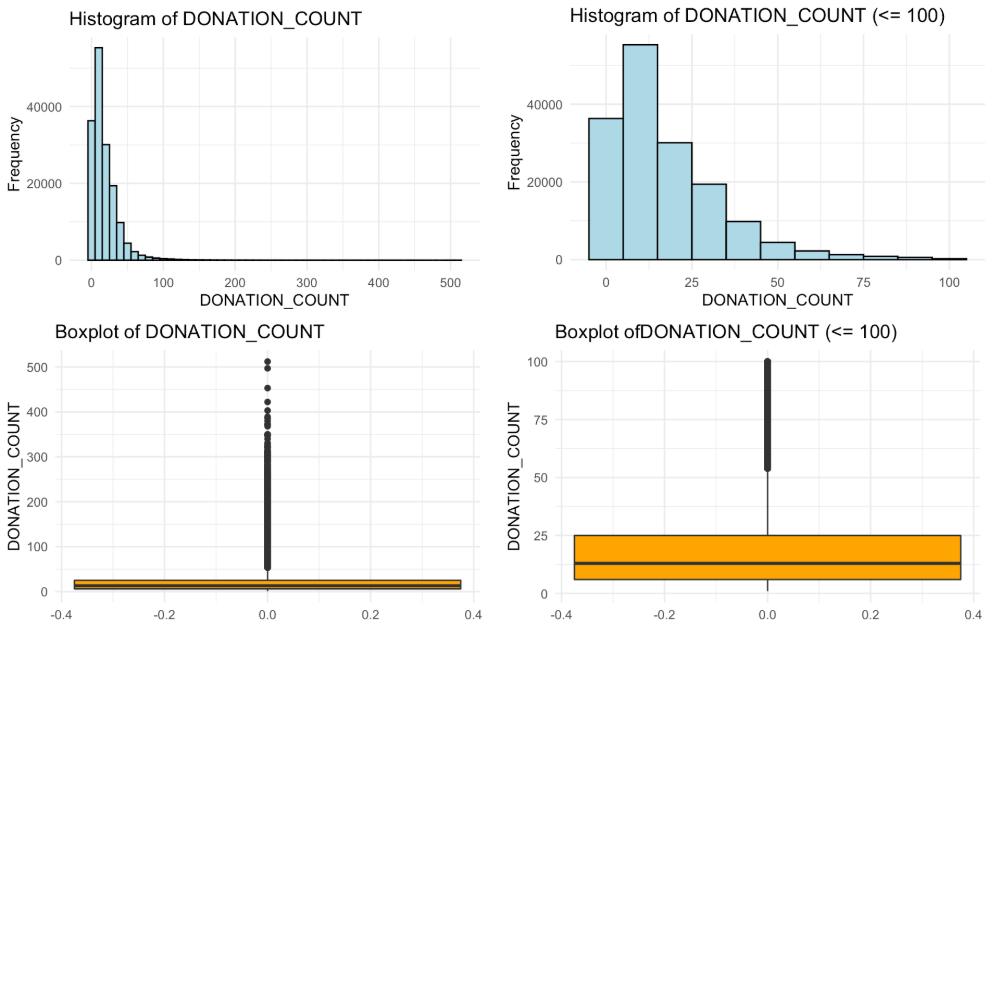
* Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*(5), 1189–1232. <https://doi.org/10.1214/aos/1013203451>
* Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
* Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer. ISBN: 978-0-387-31073-2.
* Haykin, S. (1999). *Neural Networks: A Comprehensive Foundation*. Prentice Hall. ISBN: 978-0132733502.
* Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, *45*(4), 427–437. <https://doi.org/10.1016/j.ipm.2009.03.002>
* Powers, D. M. W. (2011). Evaluation: From precision, recall, and F-measure to ROC, informedness, markedness, and correlation. *Journal of Machine Learning Technologies*, *2*(1), 37–63. <https://doi.org/10.48550/arXiv.2010.16061>
* Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. Springer. ISBN: 978-1461468486.
* James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning: With Applications in R*. Springer. ISBN: 978-1461471370.

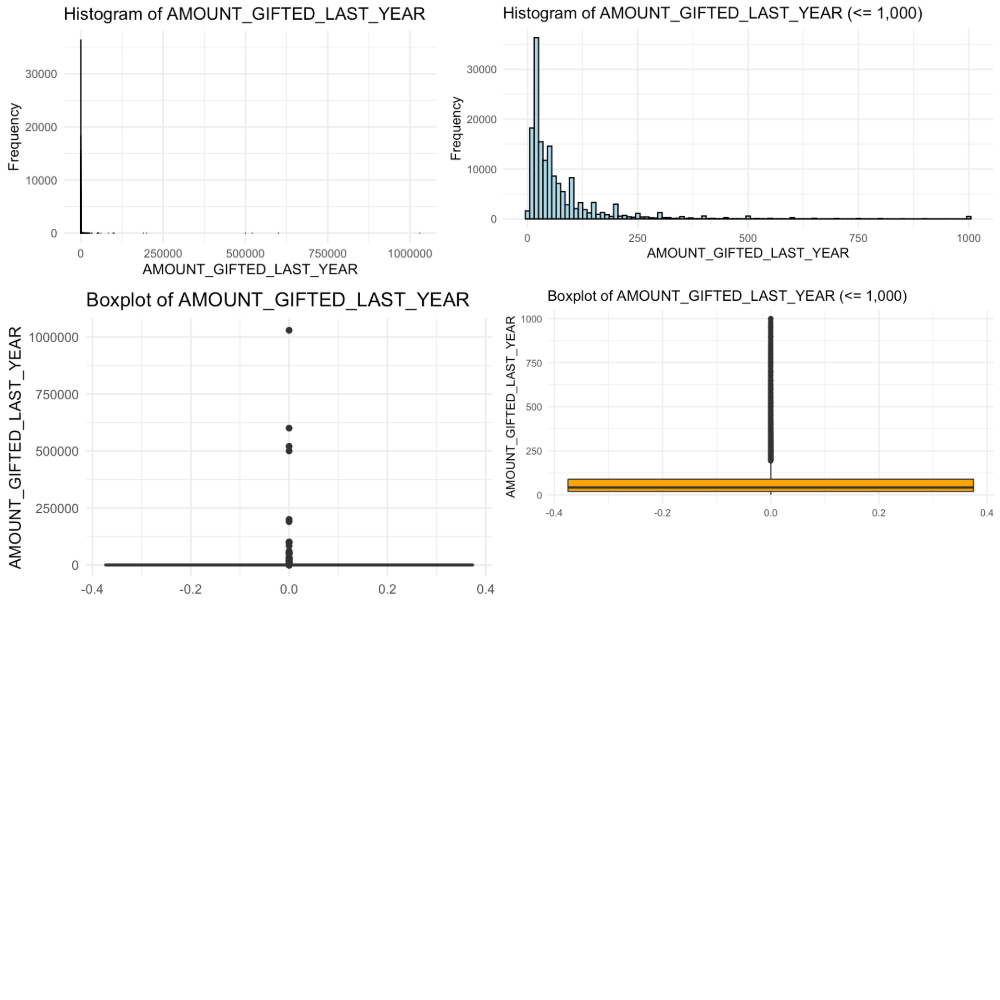
**Appendix 1**

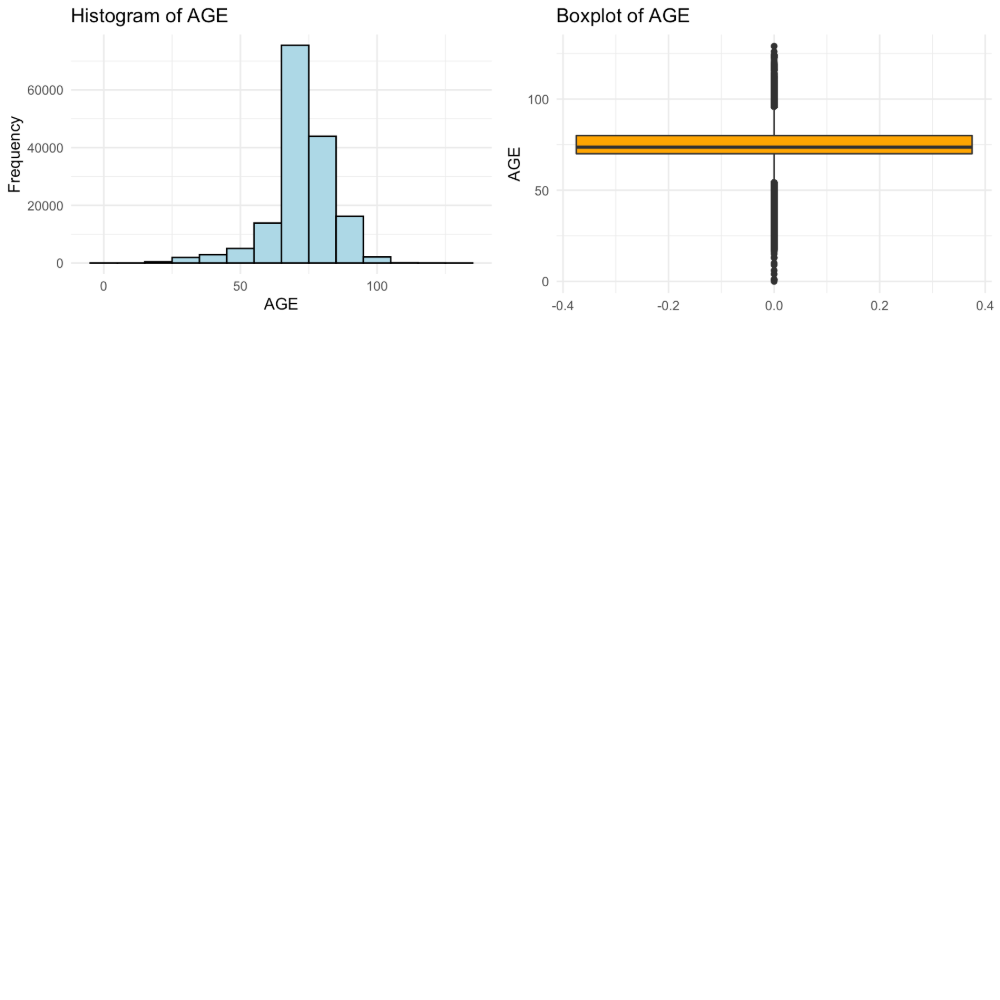
|  |  |
| --- | --- |
| **Features Used in The Generic Model** | **Features Used in The Veteran Model** |
| SUS\_FLAG | SUS\_FLAG |
| MAX\_GIFT | MAX\_GIFT |
| AVG\_GIFT | AVG\_GIFT |
| TOTAL\_GIFT | TOTAL\_GIFT |
| AMOUNT\_GIFTED\_LAST\_YEAR | AMOUNT\_GIFTED\_LAST\_YEAR |
| COUPLE | COUPLE |
| FEMALE | FEMALE |
| MALE | MALE |
| STATE | STATE |
| AGE | AGE |
| DISTINCT\_APPEAL\_COUNT | DISTINCT\_APPEAL\_COUNT |
| AVG\_DAYS\_ACROSS\_APPEALCODES | AVG\_DAYS\_ACROSS\_APPEALCODES |
| MOST\_COMMON\_GIFT\_CHANNEL\_PRO | MOST\_COMMON\_GIFT\_CHANNEL\_PRO |
| RESPONSE\_RATE | RESPONSE\_RATE |
| TOTAL\_AMOUNT\_GIFTED\_PRO | TOTAL\_AMOUNT\_GIFTED\_PRO |
| DONATION\_COUNT | DONATION\_COUNT |
| ONE\_TIME\_FLAG | ONE\_TIME\_FLAG |
| AGE\_UNDER50 | AGE\_UNDER50 |
| AGE\_50\_59 | AGE\_50\_59 |
| AGE\_60\_69 | AGE\_60\_69 |
| AGE\_70PLUS | AGE\_70PLUS |
| AGE\_NA | AGE\_NA |
| outlier\_flag | outlier\_flag |
| TOP\_5\_SUS\_PERCENT\_FLAG | TOP\_5\_SUS\_PERCENT\_FLAG |
| ORG |  |
| MEMBER |  |

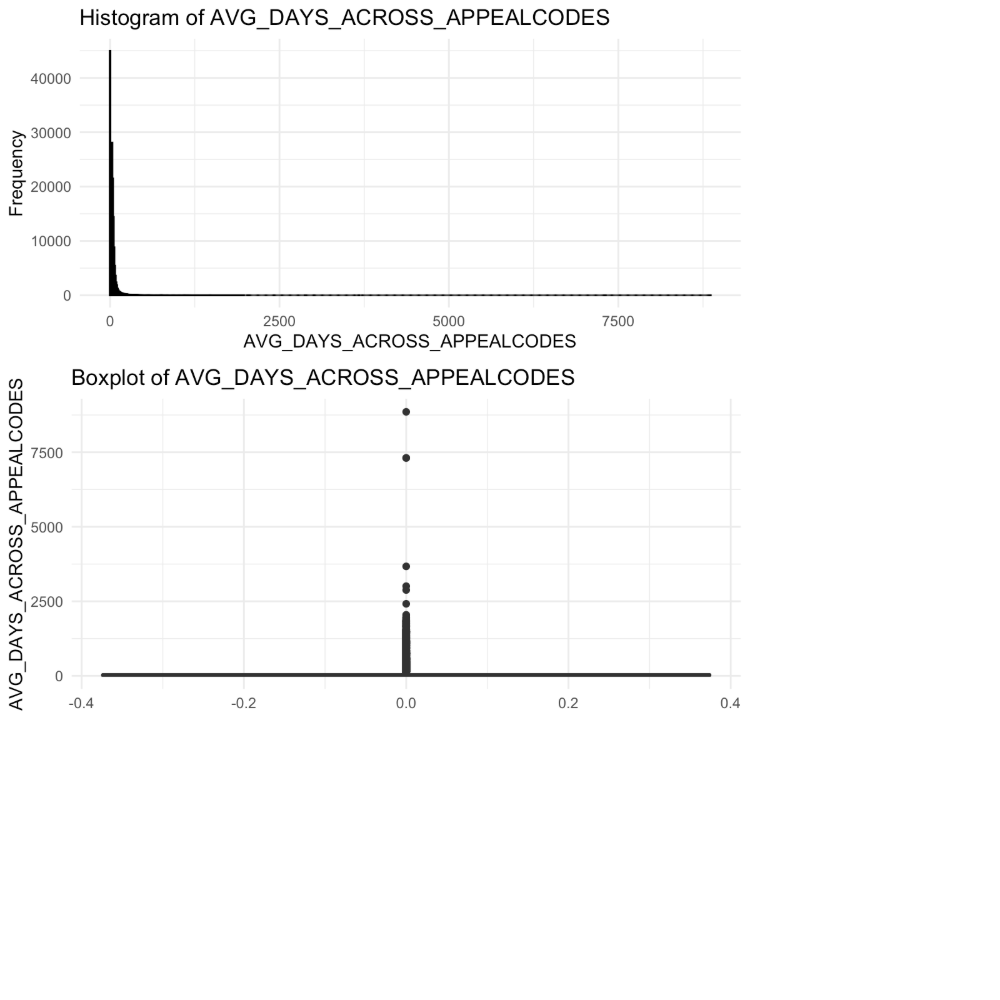
**Appendix 2**

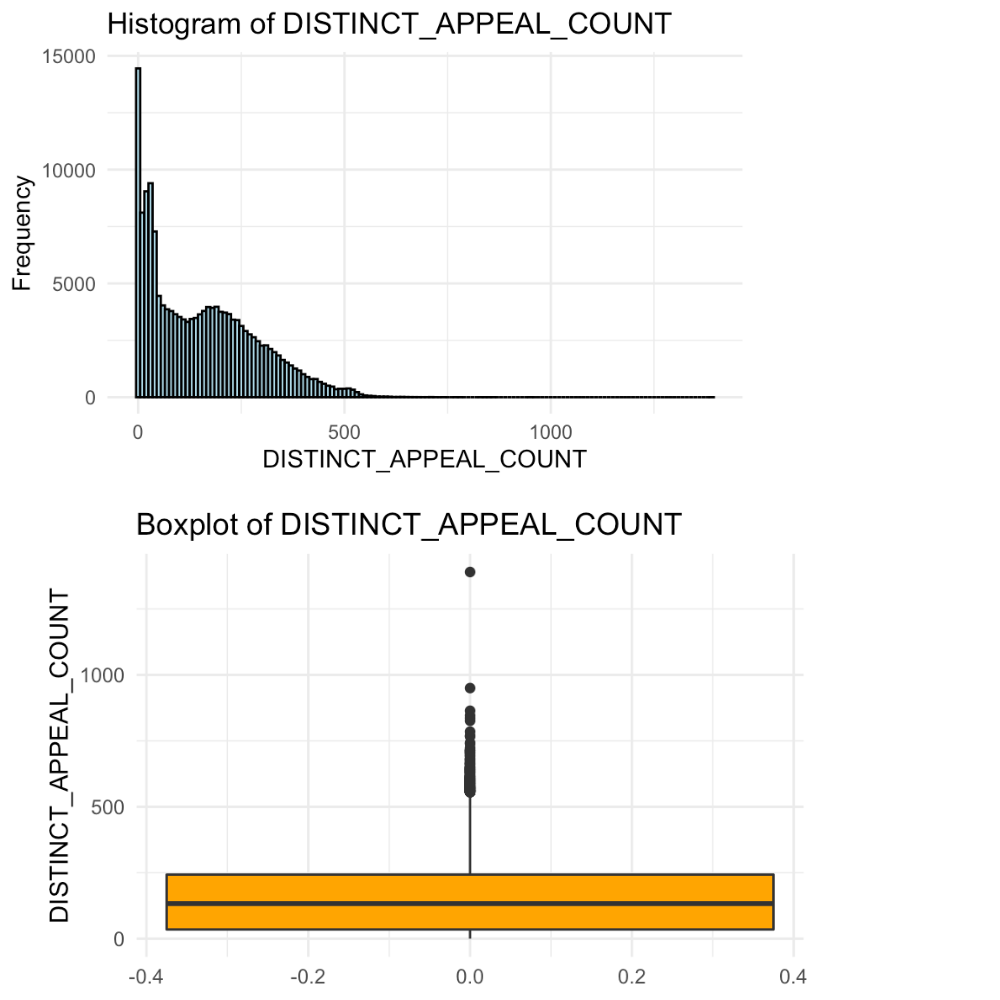


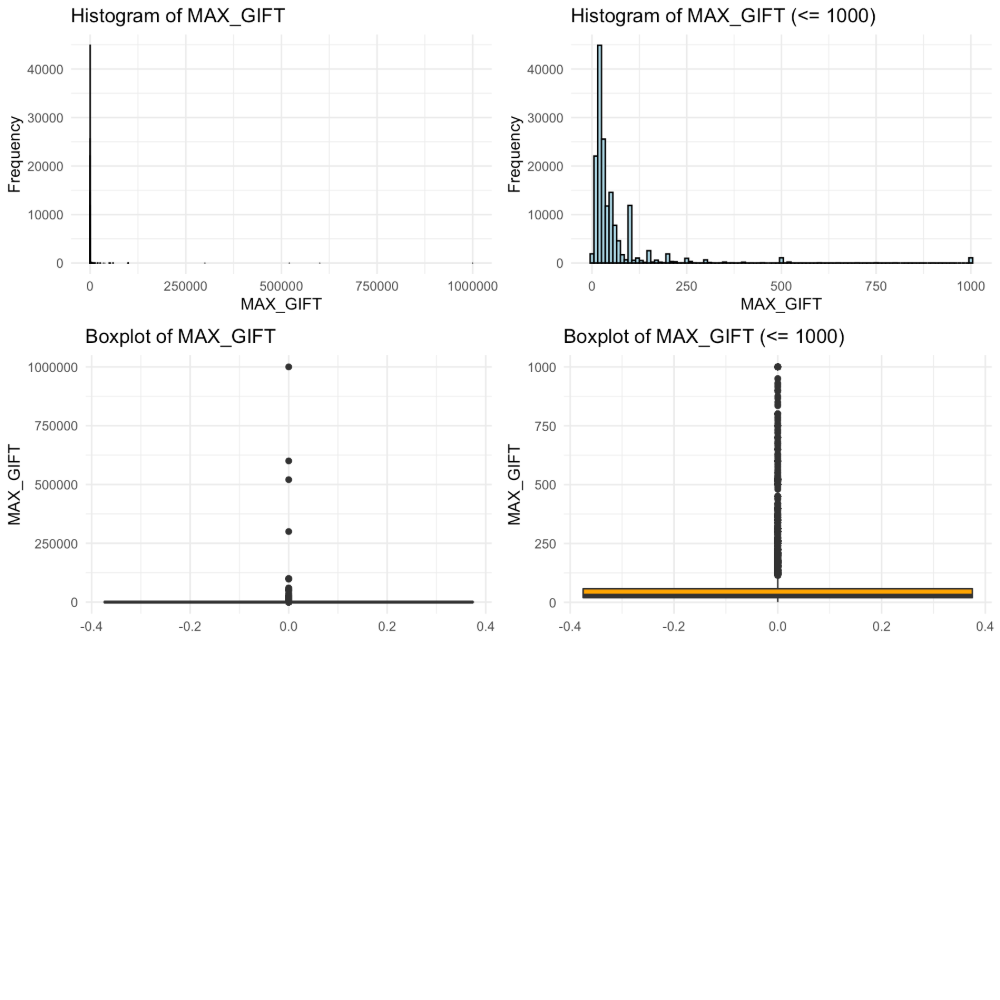


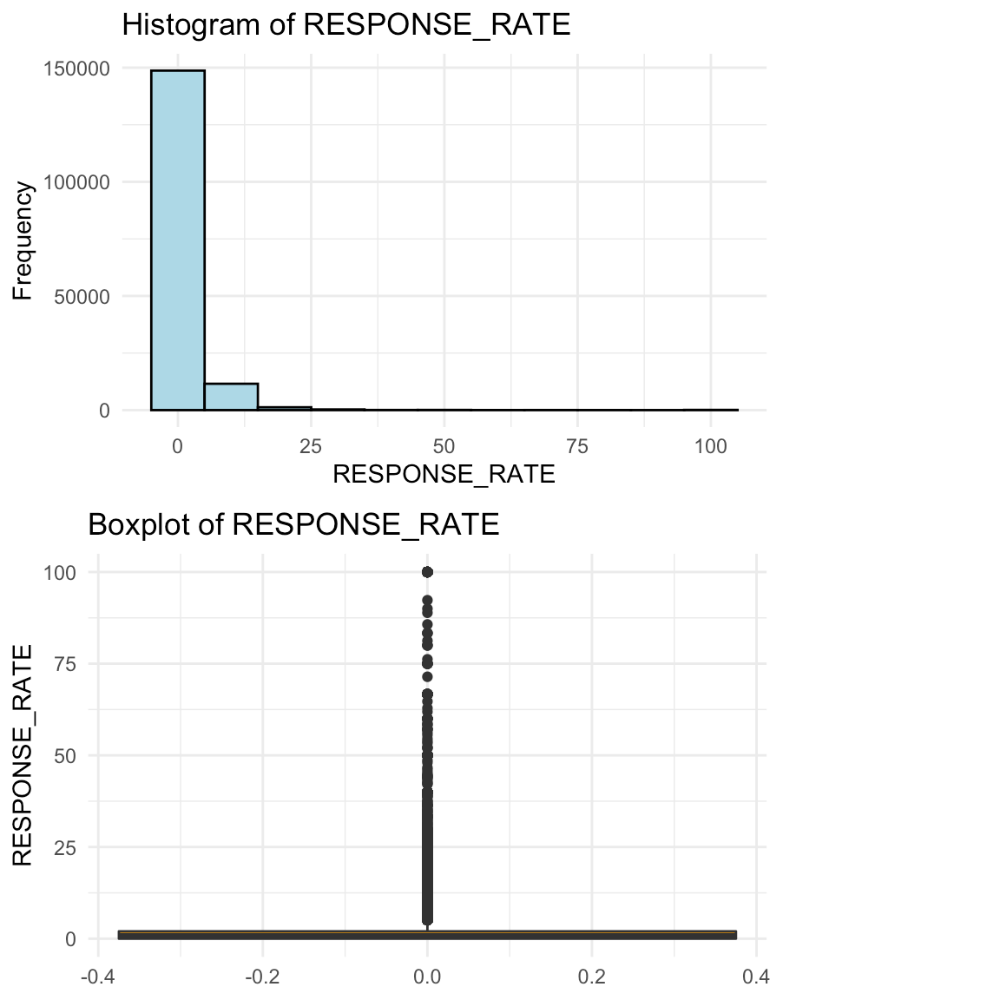


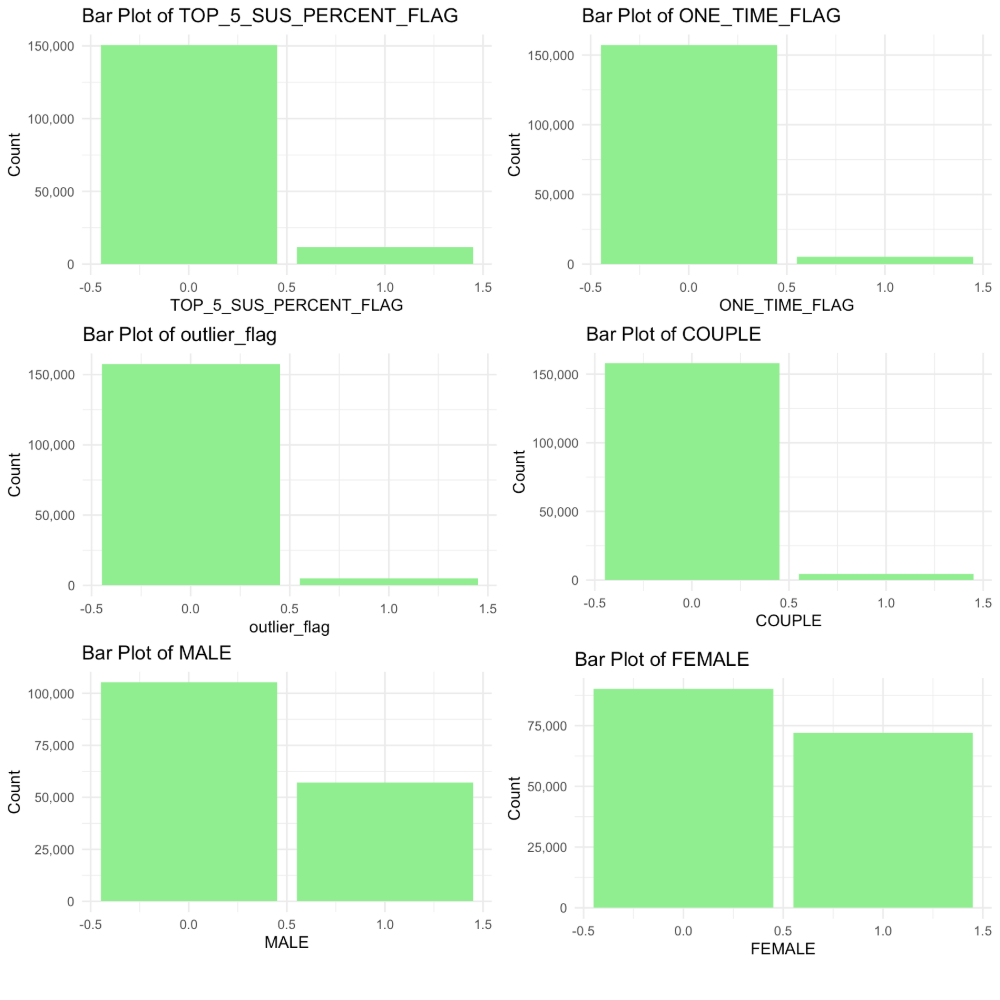
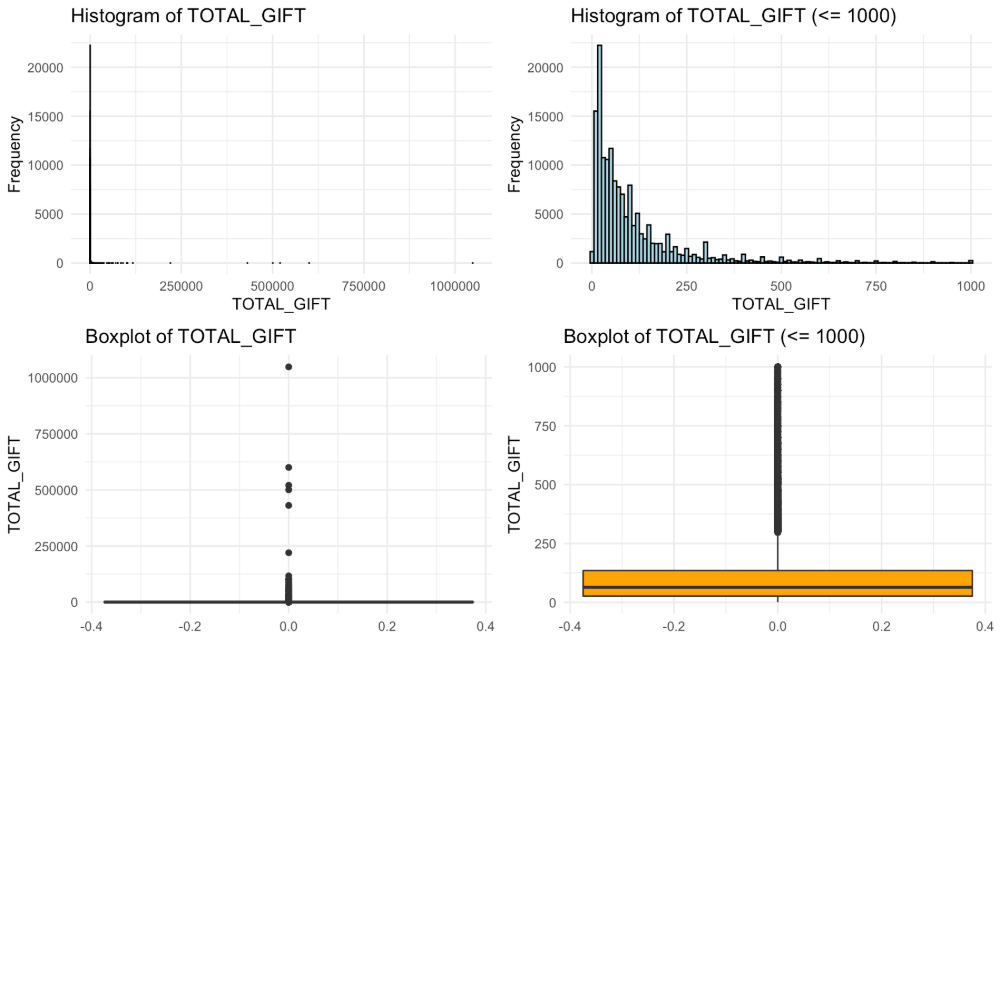






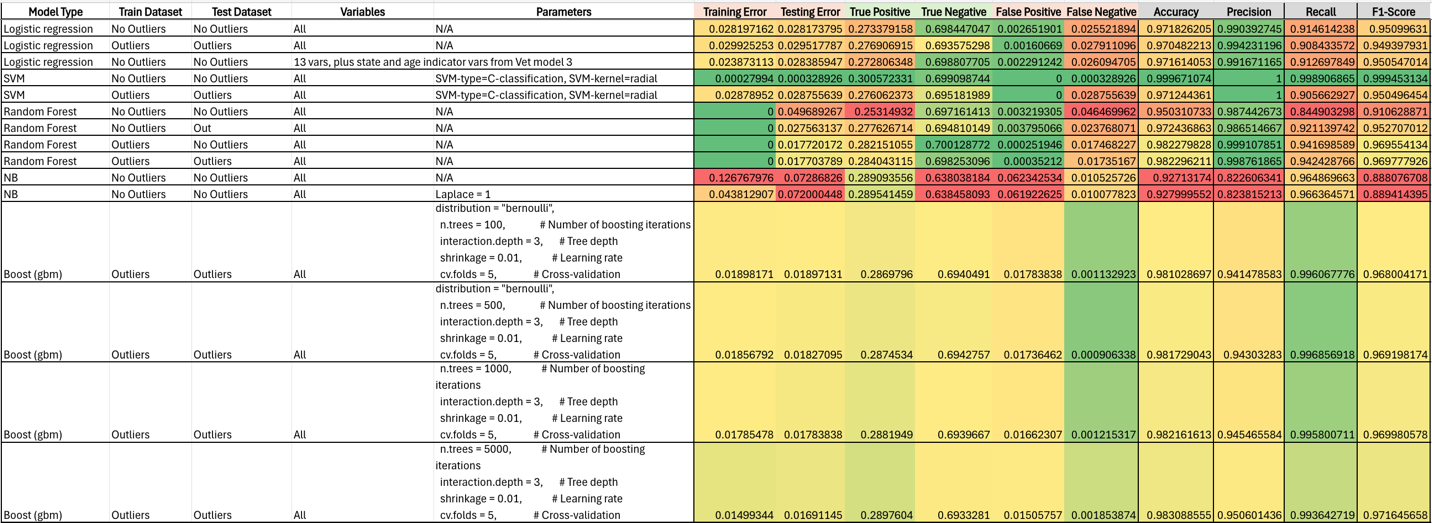


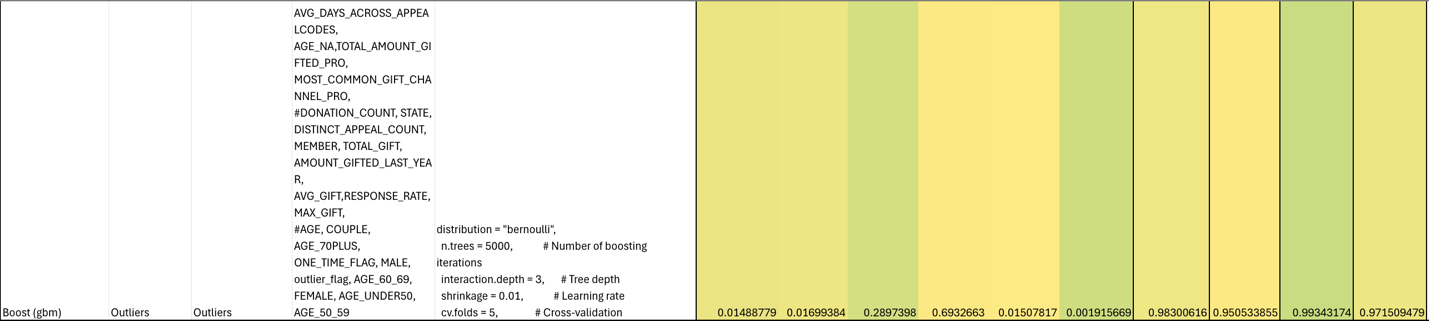


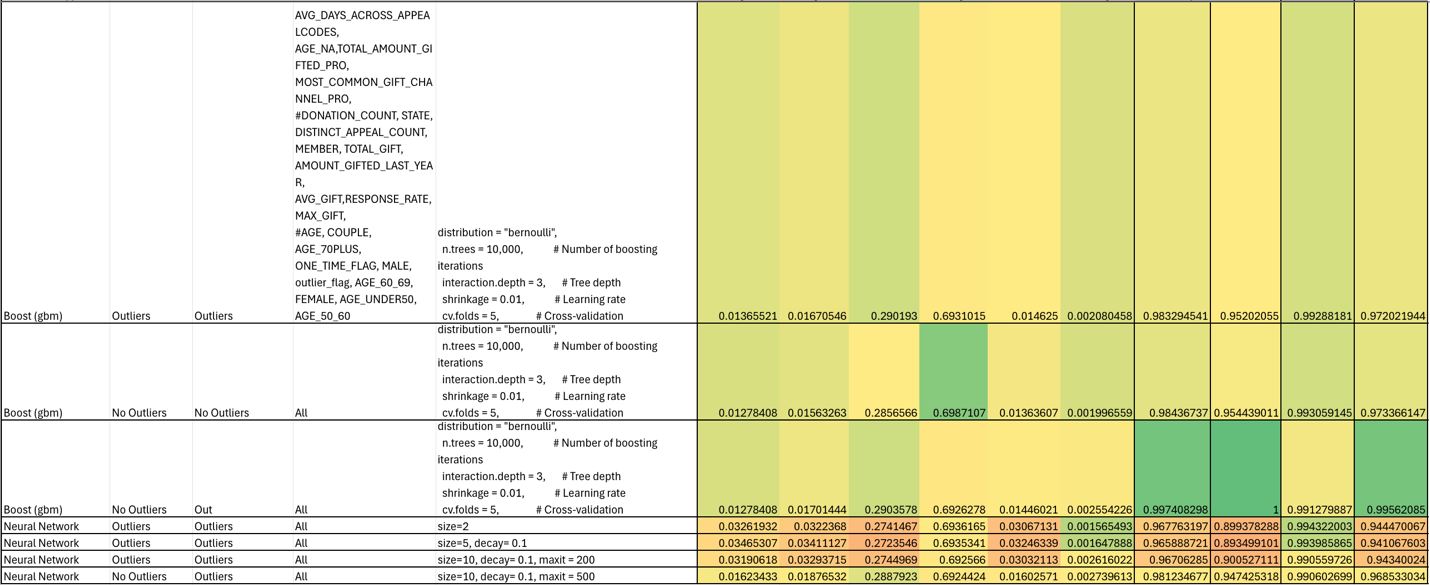


**Appendix 3**

**Generic Model Performance**







**Veteran Model Performance**

