

## **Improving Customer Targeting Through Predictive Modeling**

Dat 640: Predictive Analytics

Instructor: Jeffrey Grover

Author: Rachel Goldsbury

March 9th, 2025

## **Improving Customer Targeting Through Predictive Modeling**

Predictive analytics has become a vital tool for organizations seeking to optimize decision-making and improve business strategies. By leveraging historical data, predictive models help identify patterns and trends, allowing companies to anticipate customer behaviors and make data-driven decisions. In the insurance industry, where customer acquisition and retention present ongoing challenges, predictive modeling plays a crucial role in refining marketing strategies and reducing costs.

This report focuses on developing and evaluating predictive models to enhance customer acquisition strategies for The Insurance Company (TIC). TIC, a leading provider of insurance products, aims to improve its ability to identify potential customers for mobile home insurance policies. Traditional marketing efforts often lead to high acquisition costs and low conversion rates due to inefficient targeting. By implementing predictive analytics, TIC can refine customer segmentation and enhance the effectiveness of its marketing campaigns.

To guide this study, the research question posed is: "Can predictive modeling accurately identify potential customers who are most likely to purchase mobile home insurance, thereby improving TIC's marketing efficiency and reducing acquisition costs?" This analysis will begin with an exploration of the dataset, which consists of demographic and socio-economic attributes of customers. These variables provide valuable insights into consumer behavior, allowing for the identification of key factors that influence insurance purchasing decisions. Following this, predictive algorithms will be selected and evaluated, with Logistic Regression and Random Forest being the primary models tested. These methods will be compared to determine which model offers the best balance between accuracy, interpretability, and practicality for deployment. Once the optimal model is identified, strategies for model optimization and deployment will be

explored to ensure its successful integration into TIC's business operations.

Through this project, the goal is to conduct a thorough analysis of TIC's customer data to uncover meaningful patterns that can drive decision-making. The study will focus on developing and evaluating predictive models that effectively classify potential policyholders while continuously refining model performance through feedback and validation. Ultimately, this research will demonstrate how predictive analytics can improve customer targeting, enhance marketing efficiency, and provide TIC with a competitive advantage in the insurance market. By leveraging data-driven decision-making, TIC can reduce acquisition costs, optimize marketing strategies, and strengthen its position in the industry.

## **Background**

The Insurance Company (TIC) provided a range of insurance products to customers. The organization serves a broad customer base, utilizing strategies to improve customer acquisition and retention. TIC offers various insurance policies, including mobile home insurance, which is the focus of this analysis. The company's mission is to leverage data analytics to understand customer behavior, enhance marketing effectiveness, and optimize policy offerings. By utilizing predictive modeling techniques, TIC aims to refine its approach to targeting potential customers, ensuring personalized and efficient insurance solutions.

## **Problems and Limitations**

TIC is seeking to improve its customer acquisition strategies by identifying which individuals are most likely to purchase a mobile home insurance policy. Traditional marketing efforts can be inefficient, leading to high customer acquisition costs and low conversion rates. Some key challenges include: Identifying potential customers from large pools of prospects and Reducing marketing inefficiencies by targeting the right audience. It's important to gain an

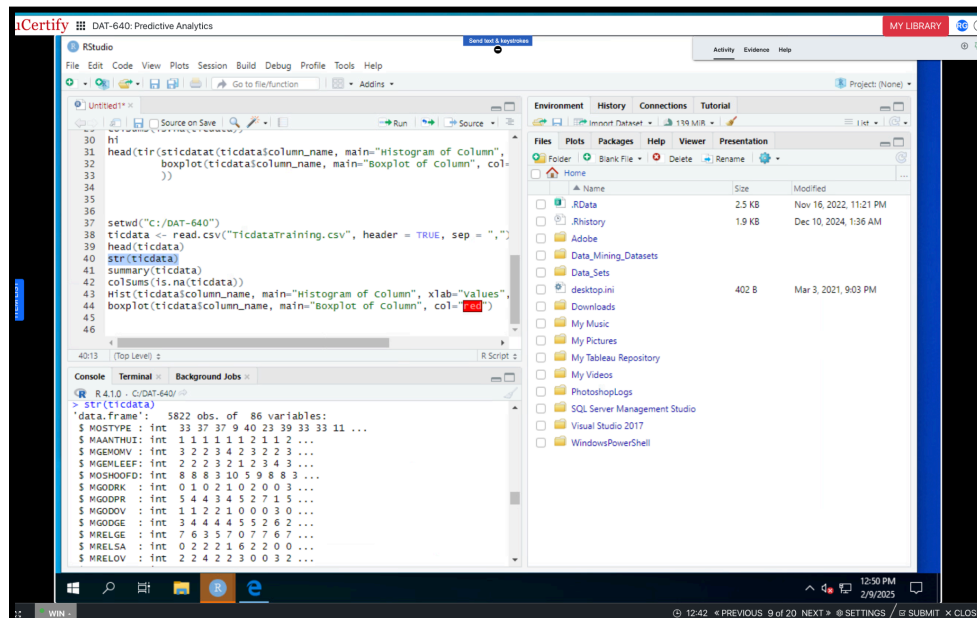
understanding of the sociodemographic factors that influence insurance purchases.

Predictive analytics plays a critical role in helping TIC overcome these challenges. By leveraging historical customer data, machine learning models can identify patterns and predict which customers are likely to purchase a policy. This helps by Improving targeting strategies for marketing campaigns and enhancing conversion rates by reaching the most relevant customers or Reducing costs associated with broad, untargeted marketing efforts. Through this predictive analytics approach, TIC can refine its business strategy, drive higher customer engagement, and gain a competitive advantage in the insurance industry.

### **Data Set Description**

The dataset used for this analysis originates from the CoIL Challenge 2000, a data mining competition that provided real-world business data for predictive modeling. The purpose of this dataset is to develop a model that predicts customer behavior regarding insurance product ownership, specifically mobile home insurance policies.

The dataset consists of 5,822 observations (rows) and 86 variables (columns), representing a range of sociodemographic and economic factors. This can be seen in *Figure 1*. These variables include household composition, income level, education status, and product ownership details. The dataset is structured in a tabular format, where each row corresponds to an individual customer, and each column represents a specific attribute related to their demographics or purchasing behavior.

*Figure 1 Overview of Dataset Structure and Variables*

To facilitate model development and evaluation, the dataset is divided into training and testing subsets. The TicdataTraining.txt file is used for training the predictive model, while TicdataTesting.txt is reserved for validation. Additionally, the ticdataTarget.txt file provides the actual target values for evaluating model performance.

## Explanation of Data Fields

The dataset consists of 86 variables, divided into sociodemographic attributes (1-43) and product ownership data (44-86). Key variables for this analysis include MOSTYPE, which classifies customers into demographic segments, and MGEMOMV, representing household size, which may influence insurance needs. Income-related factors (MINKM30 to MINK123M) and MKOOPKLA (purchasing power class) help assess financial capacity and policy ownership likelihood. Product ownership variables such as APERSAUT (car policies) and ALEVEN (life insurance policies) provide insights into financial planning and cross-selling opportunities. The

target variable, CARAVAN, indicates mobile home insurance ownership, serving as the primary outcome for predictive modeling.

### **Contribution to Organizational Goals**

These variables are essential for optimizing business strategies by enabling customer segmentation, refining targeted marketing based on income and household characteristics, and identifying cross-selling opportunities through existing policy ownership. They also support risk assessment and pricing by evaluating financial stability. Analyzing these factors helps enhance customer acquisition, retention, and overall profitability in the insurance market.

### **Summary Statistics of the Dataset**

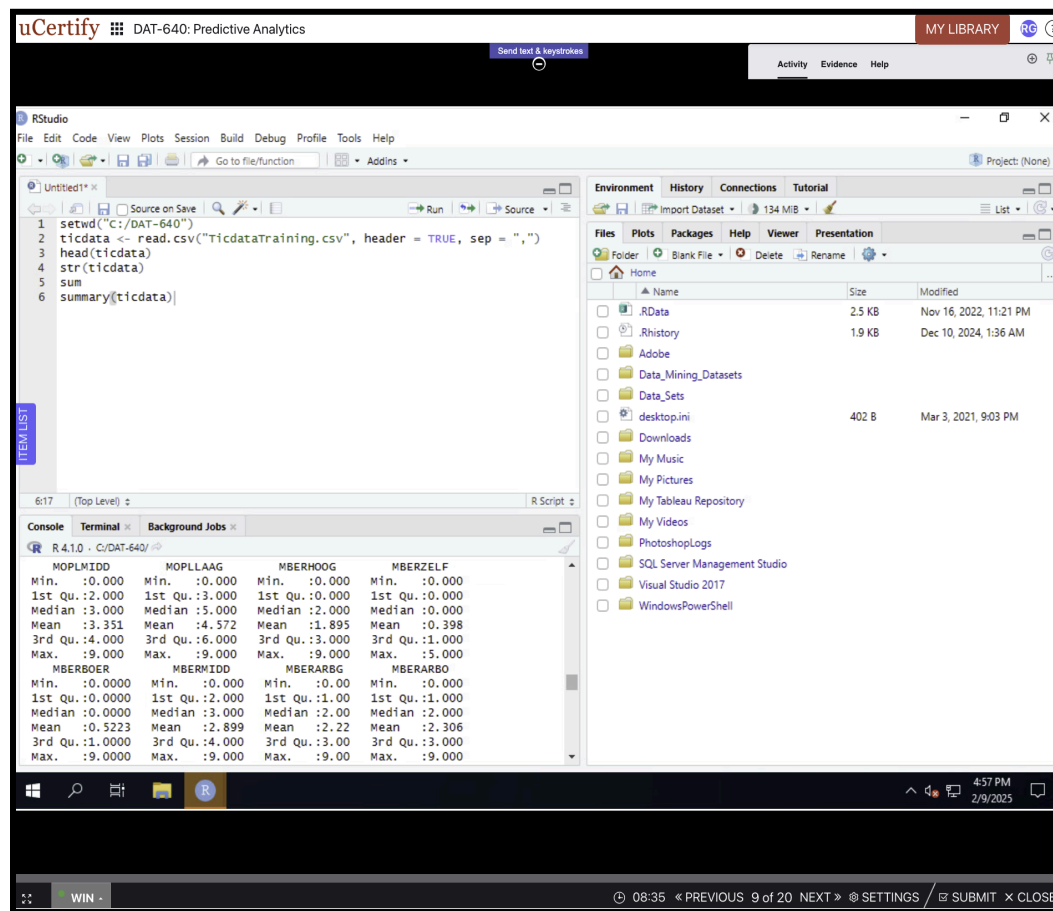
*Figure 2* below shows the summary statistics of the dataset which provides key insights into customer demographics, purchasing power, and insurance ownership. The central tendencies show that the average customer subtype (MOSTYPE) is around 24.25, with a median of 30, indicating a diverse customer base. Household size (MGEMOMV) has a mean of 2.57 and a median of 2, suggesting that most households are relatively small.

Income levels (MINK123M) vary widely, with values ranging from 0 to 9 and a low mean of 0.2027, indicating that a significant portion of customers belong to lower-income brackets. The purchasing power class (MKOOPKLA) follows a relatively balanced distribution, with a median of 4 and a mean of 4.236, spanning from 1 to 8.

For insurance related variables, car policies (APERSAUT) and life insurance policies (ALEVEN) have low means (0.01 and 0.077, respectively), with medians of 0, suggesting that most customers do not hold multiple policies. The target variable (CARAVAN) is also highly imbalanced, with a mean of 0.05977 and a maximum of 1, meaning that only a small percentage

of customers own mobile home insurance.

Figure 2: Summary Statistics of Key Variables



Overall, This dataset analysis up until this point, highlights key customer demographics, purchasing power, and insurance ownership trends, revealing an imbalanced distribution in mobile home insurance policies. Understanding these patterns is crucial for refining marketing strategies and improving customer targeting. Data exploration plays a vital role in addressing business challenges by identifying influential factors in policy ownership. Moving forward, predictive modeling will be applied to enhance decision-making and optimize customer acquisition strategies for TIC.

## **Building and Evaluating Data Analytic Models**

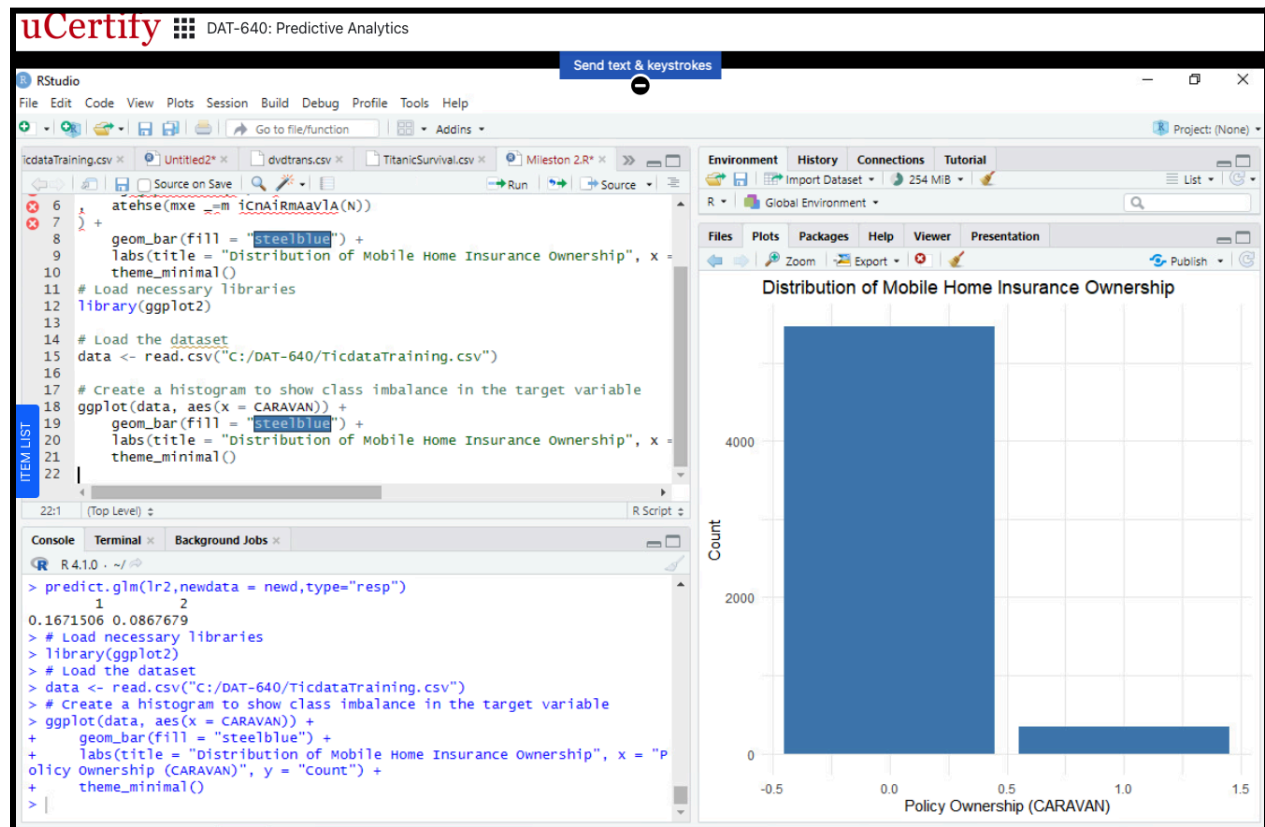
For this next section, we shift our focus on building and evaluating predictive algorithms to optimize customer acquisition strategies for The Insurance Company (TIC). The goal is to determine which customers are most likely to purchase mobile home insurance policies, improving marketing efficiency and conversion rates. By leveraging machine learning models, TIC can refine its targeting strategies and reduce acquisition costs. This paper discusses the selection of predictive algorithms, their implementation, and techniques for optimizing their performance to ensure accuracy and reliability in future applications.

One of the key challenges in building an effective predictive model is understanding the distribution of the target variable (CARAVAN), which represents mobile home insurance ownership. Since predictive models can be biased by imbalanced data, it is important to analyze how many customers in the dataset have purchased a policy compared to those who have not. The histogram below illustrates the distribution of policy ownership, highlighting the imbalance in the dataset and providing insight into the modeling challenges ahead.

The histogram shows that the vast majority of customers (left bar) have not purchased a mobile home insurance policy, with a count of nearly 5,500. In contrast, the number of policyholders (right bar) is significantly lower, estimated between 100 and 200. This severe class imbalance means that a predictive model trained on this data may favor the majority class (non-policyholders) and struggle to correctly identify potential buyers. Addressing this imbalance will be crucial for building an accurate and fair predictive model, as an unbalanced dataset can lead to poor recall for policyholders and a misleadingly high accuracy dominated by the majority class.



Figure 3 : Histogram showing the distribution of mobile home insurance ownership.



## Predictive Algorithms

### Selection Criteria for Predictive Algorithms

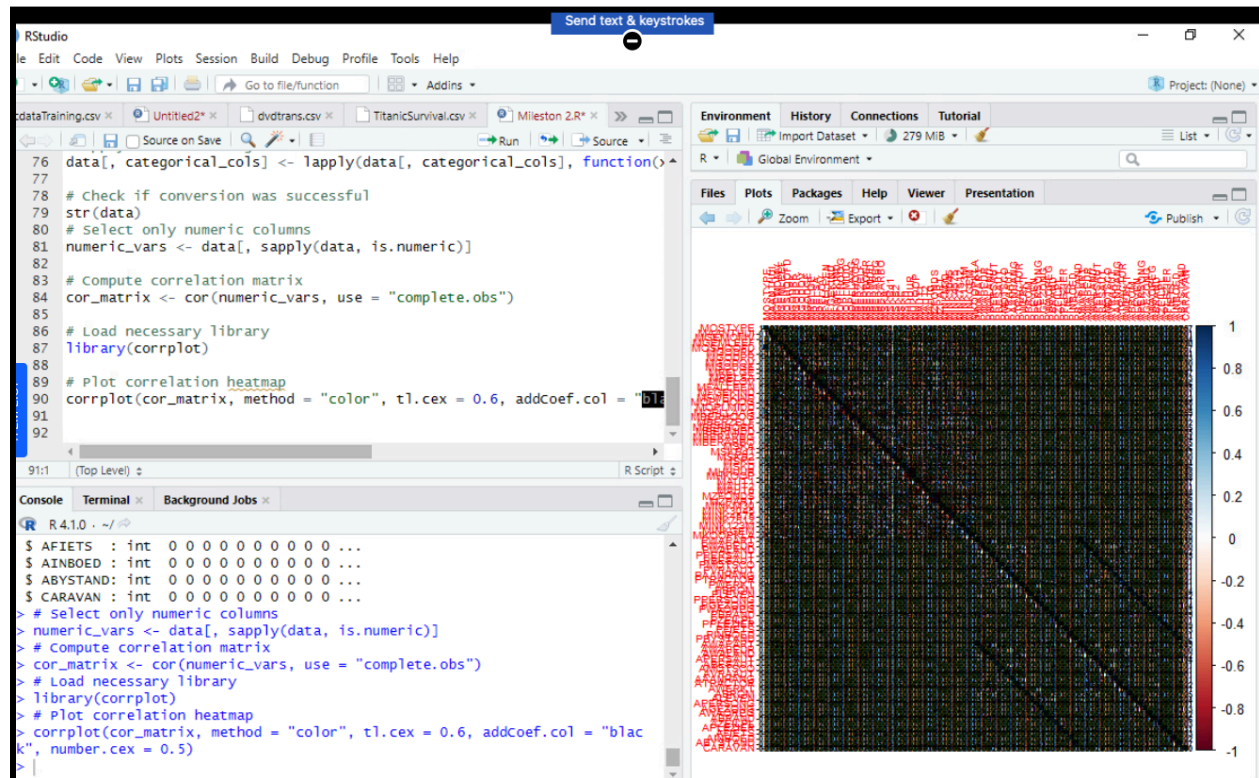
Selecting the right predictive algorithm for TIC requires evaluating data structure, relationships, and interaction specifications. The dataset consists of sociodemographic and economic factors (e.g., income, household size, purchasing behavior) that influence mobile home insurance purchases. Given that the target variable (CARAVAN) is binary (insurance policy: yes/no) and highly imbalanced, the chosen algorithm must handle class imbalance, missing values, and categorical features. Additionally, the model should provide high interpretability and strong generalization performance to ensure it is useful for real-world applications.

To better understand the relationships between independent variables and the target

variable, a correlation heatmap was generated. This visualization helps identify which features are most strongly associated with mobile home insurance ownership, as well as potential multicollinearity among predictors. The heatmap reveals that certain income-related variables, purchasing power indicators, and household characteristics show stronger correlations with the target variable, suggesting they may serve as key predictors in the model. On the other hand, some features exhibit low or negligible correlations, indicating they may contribute little to the predictive power of the model. Understanding these relationships is essential for feature selection, ensuring that only the most relevant variables are included in the final predictive models.

The heatmap reveals that income-related attributes and purchasing power indicators show the strongest correlations with mobile home insurance ownership, reinforcing their importance as key predictors. Additionally, some variables exhibit high correlation with each other, suggesting potential multicollinearity, which may require careful feature selection to prevent redundancy. Meanwhile, other demographic variables, such as certain household characteristics, display weak or negligible correlations, indicating they may have limited impact on predicting insurance ownership. These insights help refine the selection of features, ensuring the predictive model focuses on the most influential variables.

Figure 4: Heatmap



## Recommendation of Predictive Algorithm

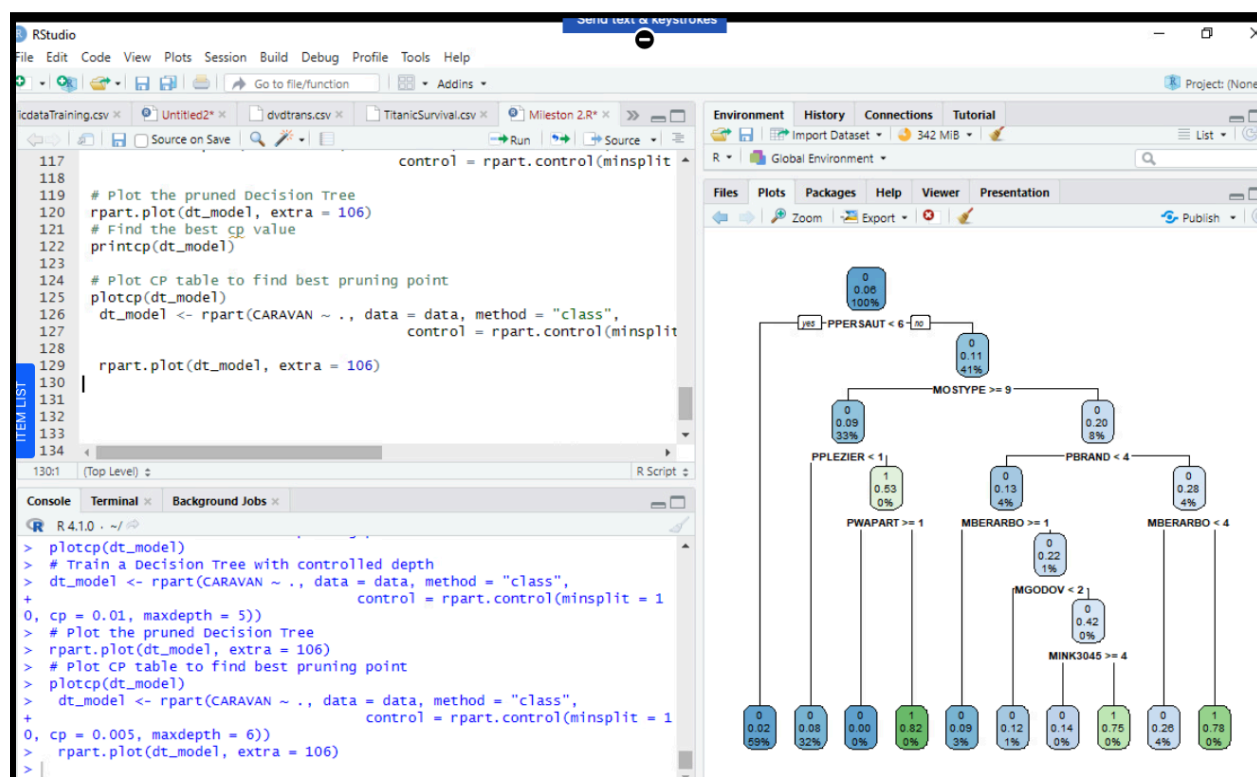
A Logistic Regression model is recommended due to its ability to handle binary classification problems while providing interpretable results. Logistic regression effectively models probability scores, allowing TIC to prioritize customers with a higher likelihood of purchasing insurance. However, to improve accuracy and capture non-linear relationships, a Random Forest classifier will also be implemented. Random Forest aggregates multiple decision trees, reducing overfitting and improving predictive performance on unseen data. Combining both models allows for a comparison of performance and reliability, ensuring an optimal solution for TIC's marketing strategy.

The Decision tree visualization reveals the step-by-step process by which the model predicts mobile home insurance ownership. The root node begins with  $PPERSAUT < 6$ , meaning

the number of car policies a customer holds is the first and most influential decision point. From there, the tree splits into two branches, with the left side (yes 0 0.02 59%) indicating customers more likely to not own mobile home insurance, while the right side (no 0 0.11 41%) continues to split based on MOSTYPE  $\geq 9$ , a demographic classification variable.

Further down the tree, the left branch (0.09 33%) splits into PWAART  $\geq 1$ , which identifies whether a customer owns apartment insurance. This factor plays a crucial role, as the model predicts that those without apartment insurance (0 0.08 32%) remain unlikely to purchase mobile home insurance, while those with it are further split based on additional financial characteristics. On the right side of the tree, PBRAND  $< 4$ , an indicator of purchasing behavior, leads to further segmentation based on MBERARBO and MGODOV, variables related to employment and education status. These final branches show that certain groups have a much higher likelihood of purchasing a mobile home insurance policy (e.g., terminal nodes with probabilities 1 0.78 0%), while others remain in the majority "no" classification.

Figure 5: Decision tree



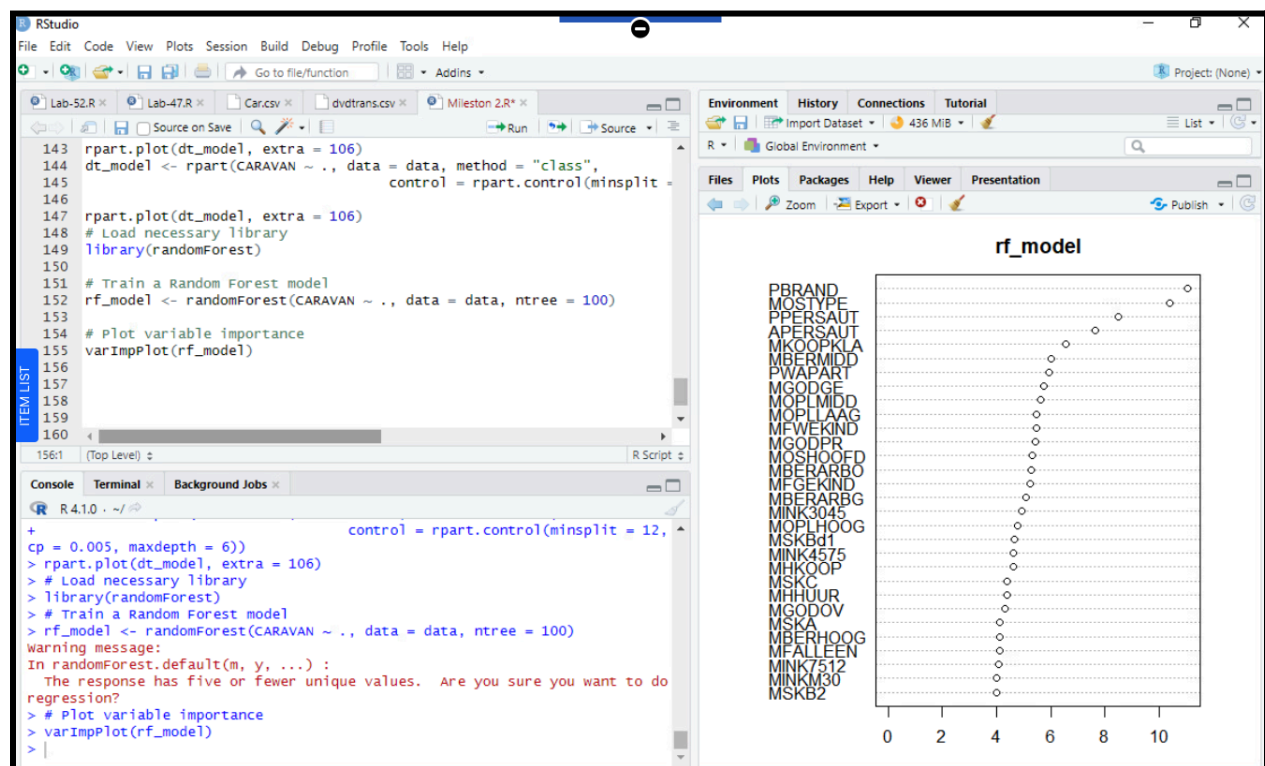
## Model Optimization

To ensure model effectiveness, Logistic Regression and Random Forest were selected for comparison. Logistic Regression is ideal for interpretable probability-based classification, while Random Forest provides higher accuracy by reducing variance and capturing complex relationships. This combination allows for a balanced evaluation of model performance in predicting mobile home insurance purchases.

The Random Forest model, figure below, provides a more comprehensive view of the most influential factors affecting mobile home insurance ownership by averaging multiple decision trees. The feature importance plot reveals that `PBRAND`, a purchasing behavior indicator, holds the highest predictive power, suggesting that past purchasing trends strongly

influence a customer's likelihood of buying insurance. MOSTYPE, which categorizes customers into demographic segments, follows closely, reinforcing the role of customer classification in predicting policy ownership. Other significant contributors include PPERSONAUT and APERSAUT, which reflect existing personal and automobile insurance policies, indicating a strong correlation between prior insurance engagement and mobile home policy purchases. Additionally, MKOOPLA, a purchasing power classification variable, plays a notable role, suggesting that financial capacity is a key determinant in insurance decisions. Below this, several other variables hold moderate influence, ranging between importance values of 4 and 6, demonstrating that while they contribute to predictions, their individual impact is less pronounced.

Figure 6: Random Forest Model



## **Reliability of Predictive Models**

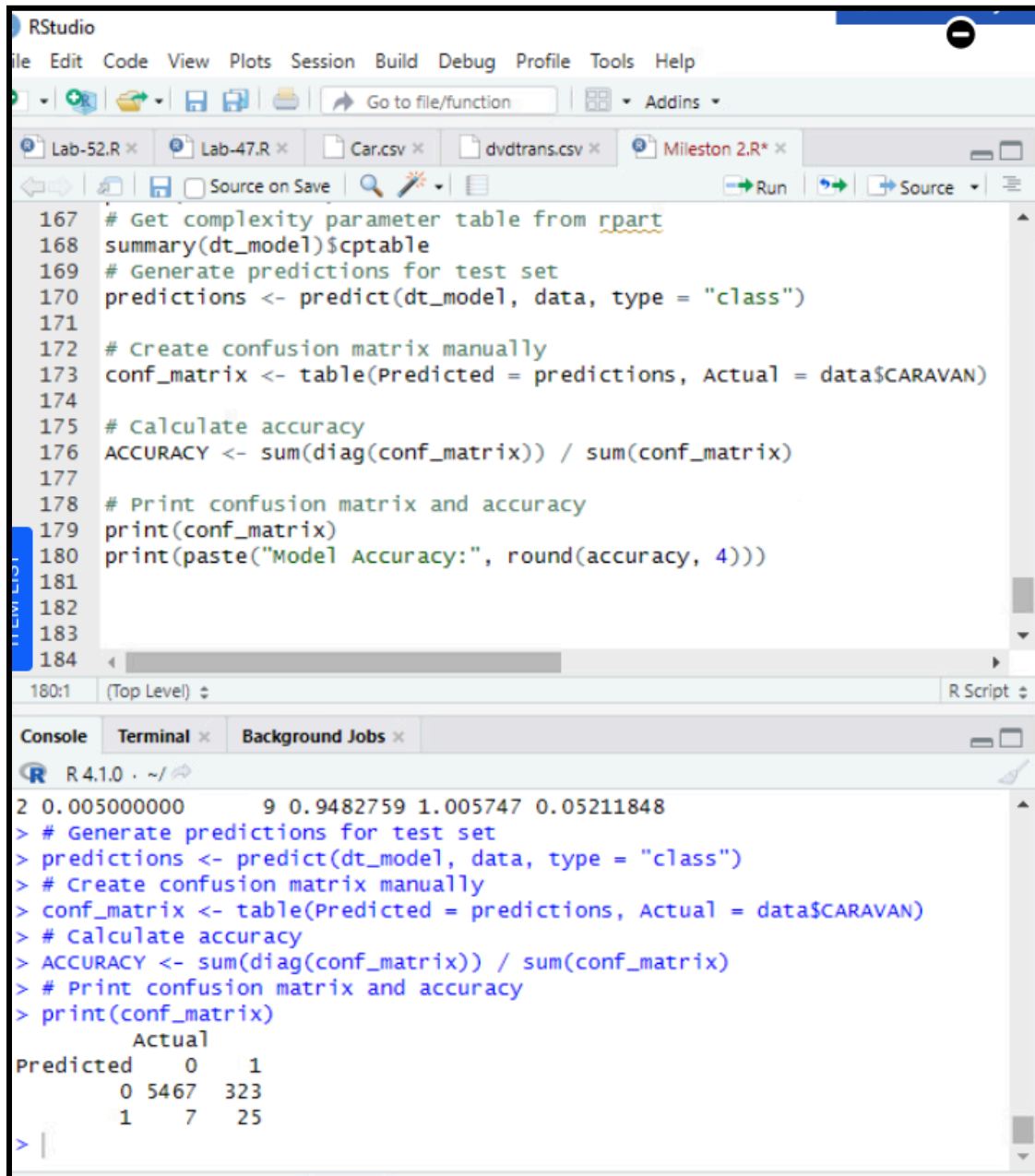
Reliability is a key factor in deploying predictive models. Logistic Regression is reliable when assumptions about linearity hold, but it may struggle with high-dimensional data or complex interactions. On the other hand, Random Forest is highly reliable, as it mitigates overfitting by averaging multiple decision trees. However, it requires more computational resources and is less interpretable. By evaluating model accuracy, precision, recall, and F1-score, TIC can determine which model provides the most dependable predictions.

To assess model performance, a confusion matrix was, displaying the distribution of true positives, false positives, true negatives, and false negatives. The confusion matrix provides insight into how well each model distinguishes between customers who are likely and unlikely to purchase mobile home insurance. This helps identify potential misclassifications and highlights whether the model is favoring one class over the other, which is particularly important given the class imbalance in the dataset.

The confusion matrix reveals how well the models classify customers as either policyholders (1) or non-policyholders (0). The majority of predictions fall into the true negative category (5,467 instances), meaning the model correctly identified customers who did not purchase a mobile home insurance policy. However, there are 323 false positives, where the model incorrectly predicted a customer would buy insurance when they did not. The false negatives (7 cases) indicate instances where actual policyholders were misclassified as non-policyholders, which is a critical error in terms of missed marketing opportunities. Finally, the true positives (25 instances) show correctly predicted policyholders, though this number is relatively low due to the dataset's class imbalance. This imbalance suggests that while the model

is highly accurate overall, it may need further optimization to improve recall and better capture the minority class.

Figure 7: Confusion Matrix



```

167 # Get complexity parameter table from rpart
168 summary(dt_model)$cptable
169 # Generate predictions for test set
170 predictions <- predict(dt_model, data, type = "class")
171
172 # Create confusion matrix manually
173 conf_matrix <- table(Predicted = predictions, Actual = data$CARAVAN)
174
175 # Calculate accuracy
176 ACCURACY <- sum(diag(conf_matrix)) / sum(conf_matrix)
177
178 # Print confusion matrix and accuracy
179 print(conf_matrix)
180 print(paste("Model Accuracy:", round(accuracy, 4)))
181
182
183
184

```

180:1 (Top Level) R Script

Console Terminal Background Jobs

R 4.1.0 ~/

```

2 0.005000000    9 0.9482759 1.005747 0.05211848
> # Generate predictions for test set
> predictions <- predict(dt_model, data, type = "class")
> # Create confusion matrix manually
> conf_matrix <- table(Predicted = predictions, Actual = data$CARAVAN)
> # Calculate accuracy
> ACCURACY <- sum(diag(conf_matrix)) / sum(conf_matrix)
> # Print confusion matrix and accuracy
> print(conf_matrix)
      Actual
Predicted  0   1
      0 5467 323
      1   7  25
>

```



Beyond overall accuracy, precision and recall are crucial in evaluating predictive reliability. Precision, which measures the proportion of predicted positive cases that were actually correct, was 0.7812, indicating that when the model predicted a customer would purchase insurance, it was correct about 78.12% of the time. However, recall, which measures the model's ability to identify all actual policyholders, was 0.0712, meaning it correctly identified only 7.12% of true insurance buyers. This imbalance suggests that the model is highly precise but overly cautious, capturing only a small subset of actual policyholders while missing many potential buyers. A higher recall but lower precision could lead to too many false positives, resulting in wasted marketing efforts. The precision-recall figures below illustrate these trade-offs for both models, helping TIC make informed decisions about model effectiveness.

*Figure 8: Precision and Recall Figures Here*

```
> # Print results
> print(paste("Precision:", round(precision, 4)))
[1] "Precision: 0.7812"
> print(paste("Recall:", round(recall, 4)))
[1] "Recall: 0.0718"
> print(paste("F1 score:", round(f1_score, 4)))
```

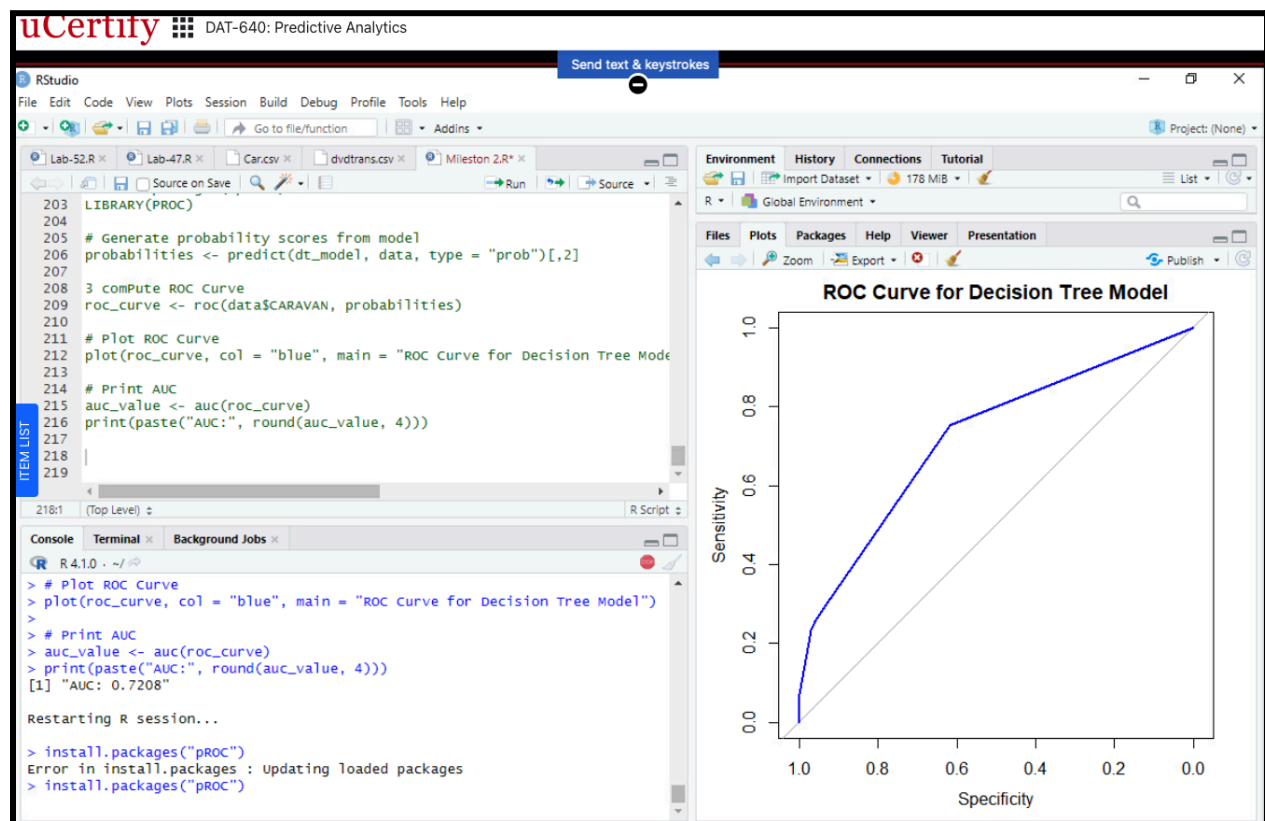
### **Continuous Feedback and Model Evaluation**

To maintain model performance over time, a continuous feedback system will be implemented to ensure the predictive models remain accurate and adaptive to changing customer behaviors. This process will include periodic model retraining with updated customer data, allowing the model to learn from new patterns and trends. Additionally, prediction accuracy will be closely monitored using key performance metrics such as ROC-AUC, precision-recall curves, and confusion matrices, ensuring that the model continues to make reliable classifications. To

further enhance efficiency, hyperparameter tuning through grid search or Bayesian optimization will be conducted, optimizing the model's parameters to improve performance. Lastly, regular data validation will be performed to detect and address shifts in customer demographics or purchasing behaviors, ensuring that the model remains aligned with real-world trends. By implementing this continuous feedback loop, TIC can maintain a high-performing predictive model that evolves alongside its customer base.

The ROC curve provides a visual representation of the model's ability to distinguish between policyholders and non-policyholders at different classification thresholds. The steep initial jump from 0.0 to 0.2 suggests that the model is quickly capturing a portion of true positives, meaning it correctly identifies some insurance buyers early on. This is followed by a rapid climb to nearly 0.8 sensitivity, indicating that the model performs well in detecting true positives while balancing false positive rates. However, as the curve progresses, the rate of improvement slows, suggesting that beyond a certain threshold, increasing sensitivity comes at the cost of a higher false positive rate. The shape of the curve highlights the trade-off between sensitivity and specificity, demonstrating the model's effectiveness in differentiating between the two classes.

Figure 9: ROC Curve to measure classification performance.



## Model Evaluation and Deployment Strategy

Through predictive modeling, TIC can enhance customer acquisition strategies by identifying individuals most likely to purchase mobile home insurance policies. Logistic Regression offers interpretability, while Random Forest improves accuracy by leveraging multiple decision trees. By optimizing models through continuous feedback and performance evaluation, TIC can ensure long-term reliability and efficiency in its predictive analytics framework. This data-driven approach will enable TIC to reduce costs, improve targeting, and gain a competitive advantage in the insurance market.

The evaluation of model reliability highlighted key trade-offs between precision and

recall. While precision was high (78.12%), indicating that when the model predicted a policy purchase, it was usually correct, recall was low (7.12%), meaning that many actual policyholders were missed. The confusion matrix and ROC curve provided further insights into classification performance, illustrating the balance between sensitivity and specificity. These results suggest that further optimization is necessary to enhance predictive accuracy.

Beyond model refinement, successful deployment is essential for ensuring that TIC can effectively integrate predictive analytics into its customer acquisition strategy. By embedding the trained model into TIC's Customer Relationship Management (CRM) system, the company can automate lead scoring and improve marketing efficiency. Deploying the model using Predictive Model Markup Language (PMML) ensures interoperability across analytical platforms, making it easier to scale and adapt the model as business needs evolve. Additionally, implementing a continuous feedback loop will allow for periodic model retraining, ensuring that predictions remain accurate as customer behaviors and market conditions shift.

### **Pilot Plan and Model Reproducibility**

For the pilot phase, TIC will implement the Logistic Regression model due to its simplicity, efficiency, and ease of interpretation. While Random Forest offers higher accuracy, it requires more computational power and tuning. Logistic Regression provides faster deployment, making it ideal for an initial test. Its probability-based predictions help mitigate the impact of TIC's imbalanced dataset, improving customer targeting while maintaining interpretability for business stakeholders. This ensures the model can be quickly deployed, evaluated, and refined to support data-driven marketing decisions.

To ensure reproducibility, the model development process follows a structured approach. Data preprocessing involved handling missing values with median imputation, converting the CARAVAN variable into a factor for binary classification, and removing highly correlated features to reduce redundancy. The dataset was then split into 80% training and 20% testing, with the model trained using a Generalized Linear Model (GLM). 10-fold cross-validation was used to improve reliability, ensuring the model generalizes well to new data.

For evaluation, the model was assessed using accuracy, precision, recall, F1-score, and AUC-ROC. A confusion matrix identified classification performance, while precision and recall provided insights into false positive and false negative rates. The ROC curve further illustrated the trade-off between sensitivity and specificity, supporting decision threshold adjustments.

To maintain documentation standards, the model workflow was structured using R Markdown, with detailed comments in the R script for clarity. Version control was implemented using GitHub, allowing for iterative improvements and collaboration. These best practices ensure the model can be easily replicated, modified, and optimized as new data becomes available.

### **Predictive Analytics Strategy**

TIC will first test the model in a controlled environment using historical customer data, rather than deploying it directly in a live campaign. This retrospective analysis will validate the model's ability to predict high-probability policyholders before expanding to real-world marketing efforts.

A sample of 1,000 customers will be selected, ensuring a mix of past policyholders and non-policyholders. The model's predictions will be compared to actual outcomes to measure classification accuracy. If results confirm strong predictive power, TIC will move forward with limited real-world deployment, using the model to prioritize marketing outreach.

Performance tracking will focus on precision, recall, accuracy, F1-score, and AUC-ROC to validate classification reliability. In a later phase, TIC will measure conversion rates (percentage of predicted leads who purchase a policy) and marketing cost reductions by comparing acquisition expenses before and after predictive analytics integration.

Post-pilot, the model will be refined based on findings. Adjustments may include recalibrating decision thresholds, refining feature selection, or retraining with additional data. Once the model consistently demonstrates strong performance, it will be fully integrated into TIC's customer acquisition workflow, driving optimized lead targeting and reduced marketing inefficiencies.

### **Communicating Value to Stakeholders**

To gain stakeholder support, TIC's leadership and marketing teams must clearly understand the business impact of predictive analytics. This model allows for smarter customer acquisition, reducing marketing waste while improving lead quality. Instead of broad, inefficient outreach, TIC can focus efforts on high-probability customers, leading to higher conversion rates and lower acquisition costs.

Predictive analytics also drives cost efficiency, ensuring marketing resources are allocated where they generate the most value. By prioritizing leads with the greatest likelihood of conversion, TIC can reduce advertising expenditures and improve return on investment (ROI).

To present findings to non-technical stakeholders, results will be shared in a visually engaging format. A slide deck presentation will highlight key insights, including predictive modeling benefits, business impact, and measurable improvements in customer acquisition. Charts, infographics, and an executive summary will help leadership quickly grasp the model's value. By aligning predictive analytics with TIC's strategic goals, this presentation will support

informed decision-making and encourage seamless adoption of data-driven marketing strategies.

## **Conclusion**

The implementation of predictive analytics at TIC has demonstrated the potential for optimizing customer acquisition strategies through data-driven decision-making. By leveraging Logistic Regression, TIC can efficiently identify potential mobile home insurance policyholders, improving marketing precision while reducing costs. The model's interpretability ensures transparency, fostering trust among stakeholders, while its probability-based predictions allow for strategic threshold adjustments to balance precision and recall. Through pilot testing, continuous optimization, and structured deployment, TIC can refine its customer segmentation approach, increasing conversion rates and enhancing marketing efficiency. As the model evolves with new data and market trends, TIC can maintain a competitive edge in the insurance industry, ensuring sustained business growth and improved customer engagement.

## References

1. *Analysis of Swiss dataset in R*. Matthias Bachfischer Blog. (2021, March 15).  
[https://bachfischer.me/posts/2021/03/analysis\\_of\\_swiss\\_dataset\\_in\\_r](https://bachfischer.me/posts/2021/03/analysis_of_swiss_dataset_in_r)
2. Author, A. the, & Tejas Tumakuru Ashok Data Scientist | AI & Machine Learning Expert | Industry Mentor He is a Data Science Excellence Award 2024 recipient with expertise in machine learning. (n.d.). *Transforming business with AI-Driven Customer Acquisition Strategies*. Data Science Council of America.  
<https://www.dasca.org/world-of-data-science/article/transforming-business-with-ai-driven-customer-acquisition-strategies>
3. GeeksforGeeks. (2023, June 5). *Bias-variance trade off - machine learning*.  
<https://www.geeksforgeeks.org/ml-bias-variance-trade-off/>
4. Google. (n.d.). *Classification: Roc and AUC | machine learning | google for developers*. Google.  
<https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc>
5. Hessing, T. (2024, February 11). *Pilot plan*. Six Sigma Study Guide.  
<https://sixsigmastudyguide.com/pilot-plan/>
6. Ibm. (2024b, December 27). *What is predictive analytics?*. IBM.  
[https://www.ibm.com/think/topics/predictive-analytics?utm\\_content=SRCWW&p1=Search&p4=43700074738575283&p5=p&p9=58700008227806587&&msclkid=f264cf8fae3316416673a84b58f60bdb&gclid=f264cf8fae3316416673a84b58f60bdb&gclsrc=3p.ds](https://www.ibm.com/think/topics/predictive-analytics?utm_content=SRCWW&p1=Search&p4=43700074738575283&p5=p&p9=58700008227806587&&msclkid=f264cf8fae3316416673a84b58f60bdb&gclid=f264cf8fae3316416673a84b58f60bdb&gclsrc=3p.ds)
7. Ibm. (2024a, December 19). *What is logistic regression?*. IBM.  
<https://www.ibm.com/think/topics/logistic-regression>



8. Meltzer, R., Rachel Meltzer Writer for The CareerFoundry Blog Rachel is the founder of MeltzerSeltzer, Rachel Meltzer Writer for The CareerFoundry Blog, Meltzer, R., Blog, W. for T. C., & Rachel is the founder of MeltzerSeltzer. (2023, August 31). *What is Random Forest? [beginner's guide + examples]*. CareerFoundry.  
<https://careerfoundry.com/en/blog/data-analytics/what-is-random-forest/>
9. Model evaluation. (n.d.).  
[https://web.archive.org/web/20170107084843/http://saedsayad.com/model\\_evaluation\\_c.htm](https://web.archive.org/web/20170107084843/http://saedsayad.com/model_evaluation_c.htm)
10. Paruchuri, V. (2012, January 18). *Improve predictive performance in R with bagging: R-bloggers*. R.  
<https://www.r-bloggers.com/2012/01/improve-predictive-performance-in-r-with-bagging/>
11. Putten, P. van der. (n.d.). *The Insurance Company (TIC) Benchmark Detailed Data Description*. Coil Challenge 2000 Report.  
<https://liacs.leidenuniv.nl/~puttenpwhvander/library/cc2000/data.html>
12. YouTube. (n.d.). *Multiple Logistic Regression*. YouTube.  
<https://www.youtube.com/watch?v=jAc2SNPTmLY>
13. YouTube. (n.d.-d). *Meta-Analysis in R for beginners*. YouTube.  
<https://www.youtube.com/watch?v=aKnsPk39vjw>
14. *12.1 - logistic regression*. 12.1 - Logistic Regression | STAT 462. (n.d.).  
<https://online.stat.psu.edu/stat462/node/207/>
15. 5-4 Milestone One. Organizational Background and Data Set .  
<https://learn.snhu.edu/d2l/le/content/1799268/viewContent/38364490/View>