**Caravan Insurance Policy Prediction Project**

Tricia M. Lang

Data Analytics, Southern New Hampshire University

DAT 610: Predictive Analytics

Dr. Taiwo Ajani

23 October 2022

**Organizational Background**

The Insurance Company (TIC) is an insurer based in Europe that offers a wide array of products. Serving over 10,000 customers, they have been in business for 72 years and are widely reputed for having great rates and customer service. They have recently looked into expanding their insurance offerings to those who have campers or recreation vehicles (RVs), with a product known as caravan insurance.

TIC would like to use their advertising budget wisely to target specific potential customers- those who likely own a caravan. To do this, they would like to build a predictive model based on data they have been able to collect, identifying possible customers. This list of customers will comprise their mailing list for the caravan insurance advertisement.

**Data Set**

Regarding the phases of the CRISP-DM model, the business understanding phase is outlined by the description of the company and what their goal is with this project. The data understanding phase is where the data is collected, explored for initial insights, evaluated for quality, and potentially divided into subsets for further investigation (Larose & Larose, 2015, p. 7). Before the data preparation phase can be initiated, the data must be thoroughly understood through the lens of the applicability to the problem that is targeted.

The data that has been supplied to build the predictive model was collected by the Dutch data mining company, Sentient Machine Research (n.d.). It includes two sets of data, a training set and test set. The training set includes information about whether the customers have caravan policies or not. The test set has a separate key to measure the predictive abilities of the model created.

There are 86 variables that are measured with 5822 rows of data in the training set. The test set has 4000 rows of data. The variables include information such as number of houses the customer owns, the size of the household, marital status, education status, employment field, social class, number of cars owned, type of health insurance- public or private, and number and type of insurance policies. All of the variables are integers with some coded for categorical data, such as age range, rather than the numeric age.

Preliminary modeling will give greater insight into the potential uses for this dataset. As pointed out by Chapman, et al (2000), the CRISP-DM model is fluid. Once modeling has begun, revisiting this phase of data understanding, as well as going back to business understanding, may occur to better define the project objective and the tools to attain it.

**Introduction**

All businesses generate data and utilizing that data as an asset and tool is how businesses can build and profit. Predictive analytics is “the process of extracting information from large daa sets in order to make predictions and estimates about future outcomes” (Larose & Larose, 2015, p. 4). A business that uses predictive analytics on the data it collects can use that information to improve performance and reduce wasteful costs.

Predictive analytics are an effective way to assist businesses with optimizing the utilization of resources. Allowing the business to focus on the target audience, resources can be used much more effectively (Marketing Evolution, 2022). According to a Rakuten Marketing survey cited by Benes (2018), “(marketers who responded) estimated they waste an average of 26% of their budgets on ineffective channels and strategies.” Identifying the most effective channels and strategies can be achieved through predictive analytics.

The data that is collected by TIC includes information about customers, including sociodemographic information and property owned. As TIC looks to expand products offered and target advertising to the individuals most likely to use those products, individuals likely to belong into that target category need to be identified. Predictive analytics will refine this list of customers to those who would potentially purchase a specific policy, and within that target audience, the means to reach those individuals can be further refined.

For TIC, the product they are looking to launch and advertise for is an insurance policy that covers campers and RVs towed behind a car. It will be important to know who has a towed caravan as those who do not have this would not purchase a policy. In addition to this, individuals who do not insure any of their property would be less likely to insure this property. Knowing who insures their property increases the likelihood of identifying the correct individuals for the target audience.

Predictive analytics also would be able to identify individuals most likely to respond to advertising in the methods chosen by TIC. Both finding the individuals and the means to reach them are cost effective benefits for the company by utilizing predictive analytics (The Data Company, 2020). Building and retaining that customer base will be profitable to TIC in the long-term. An article by Dwilson (2016) identifies which groups are most likely to respond to the mailings that TIC intends to use. Breaking the target audience down further by these categories may help with improving effective advertising.

For the individuals that do not respond to direct mailings, information from Altman (2017) may help identify an advertising strategy that would best attract those potential customers. Examples of this include digital advertising for young adults or television ads for older males. To reach female consumers, it is more effective to use print advertising within sections of magazines or newspapers with interests that are often identified as female-focused, such as fashion, home, and family.

The use of predictive analytics for TIC, and any company, in general, can improve a company’s Return on Investment (ROI). By reducing waste and identifying a target audience, predictive analytics can reduce time and money spent directing efforts towards those who have no use for the product. The cost of implementing this is significantly less than the cost of excessive and unnecessary advertising.

**Predictive Algorithms**

Deciding on a predictive algorithm is dependent on understanding the dataset and understanding the research question. The question that TIC would like to have answered is: which individuals are most likely to buy a caravan insurance policy. The dataset includes multiple datapoints on individuals including demographic, property ownership, and insurance policy holdings. Based on the information available, the predictive algorithm should create a model that can predict a greater likelihood of purchasing a caravan insurance policy.

The outcome expected is a binary “yes” or “no” for the question “is this person likely to buy a caravan insurance policy.” This narrows the type of model that would be necessary to accomplish this outcome. According to insightsoftware (2022), the top five predictive analytics models used are classification, clustering, forecast, outliers, and time series. Being that the answer sought is binary and the question is not complex, such as asking for a specific number output, the question is a classification one.

Within the grouping of models that address classification questions, there are five algorithms highlighted in an article by Sharma (2021). These include logistic regression, Naïve Bayes, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Decision Trees. Addressing which predictive algorithm is best likely to provide these results, the type of data explored is important to know. Given the dataset for TIC has a lot of variables that may or may not be related, some of these models can be ruled out.

The Naïve Bayes algorithm assumes none of the variables are related to each other (Sharma, 2021), but in the case of the dataset for TIC, it is possible some of the variables are related. For example, if one does not own a car they are not likely to have a car insurance policy. The KNN algorithm addresses similarities between data points, but when there are numerous variables it can become much more cumbersome. The SVM algorithm is often used in classification problems, but depends on the data being linear, which is not necessarily the case for the TIC data.

The two resulting options are logistic regression and decision trees for classification questions. Reviewing the pros and cons of these two models in answering a binary classification question, Lee (2016) points out that the logistic regression provides coefficients that identify which factors are more influential than others. A logistic regression will split the data along a regression line where points on one side fall into category A and the other side category B. For some data sets that are not as linear, this can be oversimplifying the results.

Lee also describes decision trees as potentially becoming overloaded with numerous decision nodes, complicating the interpretation (2016). However, the plot of decision boundaries is divided into more regions that are more specific to the outcome. This only becomes problematic when there is an overfitting of the training data, which results in less generalizable outcomes compared to a logistic regression.

Considering the needs of TIC, a decision tree model would be the most appropriate. Given the number of independent variables assessed as well as the question being a binary classification, a decision tree will be able to handle the inputs with a reliable outcome. The company does not have a need to know the probability of a person buying a caravan policy, but only whether they are or are not likely to. As pointed out by Williams (2011, p. 245), the incorporation of multiple models tends to be more effective over a single model. As such, a random forest model may be most effective and reliable to address the needs of TIC.

**Predictive Algorithm Recommendation**

Based on the research question posed by TIC: “is this person likely to buy a caravan insurance policy,” the predictive algorithm recommended is a random forest. The answer to this question is a binary “yes” or “no.” This indicates it is a classification question (insightsoftware, 2022). For models that work best with classification questions, there are five identified by Sharma (2021), though of those five some are eliminated by the ways the data are related that are not representative of data from TIC. This leaves the options of logistic regression or decision trees.

Logistic regression outputs are best used for identifying which factors have greater influence on the outcome variable, and how much of an influence (Lee, 2016). Decision trees are better equipped to handle results that are not necessarily on a linear continuum. As Lee (2016) points out, the decision boundaries for plots of outcomes in a decision tree are more specific to the outcome, rather than assumed to be linear.

For TIC, the recommended tool is a random forest. A random forest is the aggregation of numerous decision trees into one model. The incorporation of multiple models, according to Williams (2011), tends to be more effective than a single model. By using a random forest, within this one model numerous models will be used.

The random forest will be able to be tailored and pruned in a way that is best representative of the data, as well as provide feedback on which variables are important and relevant to the model. This will provide feedback to the data gathering section of the team to indicate which bits of information are important to be reported on and which are less so. Additionally, the random forest will be able to integrate new variables and reshape the model as needed should TIC provide additional information.

Using a random forest will also set the groundwork for additional data analytics that can be of use to TIC. By having the dataset available and modifying the target variable and the inputs, different questions can be answered. Reading the results summary would also provide insight into patterns for different customer bases, such as those who are likely to have caravan policies may also tend to have another trait noticed in the data that is correlated to other information. Such as those who have caravan policies may be a specific gender, age group, and income category, which may also be found to correlate to individuals who have homeowners’ insurance policies. Random forests are recommended to TIC for their reliability, validity, and potentially usable in multiple capacities.

**Model Optimization**

**Evaluation**

The predictive model used for the TIC data is a random forest model. This model is composed of a number of decision trees. The number of trees can be modified based on the evaluation of the results as being accurate and reliable. The random forest will select random variables as split points within the trees (IBM, 2021), rather than as with a decision tree where all variables are considered.

The data is divided into three groups, test, train, and validation. Each of the trees is made with a random sample of the training data known as bootstrap samples (IBM, 2021). The training data is used for building the model. The validation data subset is used for validating the data and providing the performance scores when evaluations occur.

One third of the training data subset is set aside to test the training model, this test data is the out-of-bag (OOB) sample (IBM, 2021). The OOB sample is used for validation and provides a OOB score, which is particularly useful with data sets that are small (Bhatia, 2021) when validation subsets are not feasible. The test subset is used to evaluate the model for fitness for deployment once completed.

Three ways the model will be evaluated are a confusion matrix, ROC/AUC, and scoring. The confusion matrix provides values for the model sensitivity, specificity, accuracy, and the size of the data that fall into the categories of being true or false positives or negatives. The ROC is a visualization for the predictive abilities of the model with the AUC being the quantification of performance.

The scoring of the model gives the predicted outcomes of the data points, which for TIC would be whether they would be likely to purchase a caravan policy or not. Scoring the testing data set gives an idea of how well the model performs when used on the testing data (Williams, 2010), but is not appropriate for the model development. Scoring allows the distributions of the results to be visualized as well with the use of augmenting services, such as Excel, as the output is a csv file.

**Implementation**

Scoring a model is the opportunity to apply the model to a new dataset (Williams, 2011). This can be done on the validation subset of data, test subset, or new data. By scoring data sets the validity of the model can be increased (Harris, 2020) as new information leads to potential for improvements in the model. It also allows for the most accurate model to be chosen (DataRobot, 2022). As such, this is an important part of the model building and implementation.

Once the model is created, it will be deployed in order to use the predict() function (Williams, 2011). The model will be saved and written to a csv file. At this time, it will be available to load in R and use with a novel data set. When using the test or validation data sets to evaluate the model, the results are known and the predicted values can be compared to the actual values. A model can also be deployed into a scoring engine that will give a more accessible interface for others (Maraz, 2018).

To have a model be accessible in other software, instead of direct deploying into R, the model will be exported to “an open standard format” (Williams, 2011). Using the standard language of Predictive Model Markup Language (PMML), the model will be able to be utilized with other programming languages than just R. To do this, the packages of rattle, rpart, and pmml need to be installed.

Williams (2011) outlines the process of exporting the model in PMML. The data set is saved as a dataset object using the new.env() function. The evaluation dataset is established with evalq({ data, nobs, vars, set.seed(), train, form}, theDS). The model is then built using the rpart library by establishing the model name as a new.env(parent=theDS) with the evaluation being done with evalq({ model <- rpart(formula, data)}, modelDS). At this time the PMML is created with the pmml library loaded and using the pmml() function on the rpart DS, modelDS$model.

**Expected Results**

Using the scoring engine, the output of results is expected to be a spreadsheet of predictions for each row of data. The predictions can be categorical, as with the TIC data, for predicted to be a yes (“1”) or no (“0”), or a probability, such as a certain percent likelihood of being “1” or “0” (DataRobot, 2022). The spreadsheet can also be used for further analysis, as Williams (2011) points out. Distributions can be visualized and graphed whether in Excel or another spreadsheet program, or re-entered into Rattle.

Potential unexpected results are having all results come back as all “0” or all “1”. This may be due to an overfitting of the model where the parameters were too specific to the data set it was built on and unable to account for the small proportion incidents that may have an impact on the ability of the model to predict on novel datasets (Elkan, n.d.). Having an overfitting model results from “patterns in the training set that are not statistically reliable, and are not valid in test data” (Elkan, n.d.). Should the training dataset be too small or the parameters chosen too narrow, this is a possible outcome.

**Predictive Algorithm Technique**

The random forest will be created using the programming language R and the Rattle library within R. All code that is created and processes used, as well as data sets, will be maintained in a compendium to ensure the accessibility and checking of the work and validation of the results. The compendium will allow anyone with access to reproduce the results (Gentleman & Temple Lang, 2007). The transparency in the process will allow outcomes to be trustworthy and open to improvements from outside those who are working directly with the program.

The raw data received from TIC will be cleaned before evaluation. This will include identifying structural errors in the data, irregularities, and missing values. The handling of irregularities and missing values will be determined by the number and impact of those values (Burns, 2022). Those data points may be discarded, replaced, or imputed, depending on the circumstances, and would be decided upon in a way that has the least detrimental impact on the validity and reliability of the outcome. The handling of these events will be documented for consistency in future cleaning and understanding any variance in the data.

The cleaned data will be loaded into Rattle to perform modeling by choosing the target variable of "CARAVAN:Number of mobile home policies". The data will be partitioned into 70-30 train and test datasets, where 70% of the datapoints will be used to train the model and 30% to test. Pre-pruning will occur to remove variables that may have the potential to be used unethically. This can include the variables of religion or source of medical insurance. Customer main type may also be regarded as a questionably unethical categorization of people, with descriptors such as “successful hedonists” and “career loners.” This would be raised with the risk management department to determine if these are valid categories to incorporate or would involve potential discrimination.

An initial random forest will be created using the train data in the Model function of Rattle with two variable options and 500 trees. Once this model is created the model can be evaluated for accuracy through the Evaluate tab and creating an error matrix. The importance of the variables can also be charted to identify which variables are noted to be the most pertinent to the model.

Post-pruning occurs after the model has been run and important variables identified. Modifying the number of trees and the combinations of variables, then running the evaluation for the error matrix again, will allow the best model to be produced. Once the most accurate model is created, the variables will be identified, and the resulting model will be tested. The test data will then be run through the model to validate the results. Should the results of testing indicate the model is accurately reflecting the anticipated outcome, the model will be saved for future use. Should the results indicate a problematic model, the process will restart with ensuring the data is thoroughly cleaned before modeling and adjustment of variables.

**Pilot Plan**

Based on the needs of TIC, the pilot plan for implementing a predictive analytics strategy will have five main parts, as outlined by Newton (2019). The predictive analytics plan will be presented to TIC’s executive leadership to gain support in the implementation of the plan. With top-down support and enthusiasm for the changes to be implemented the acceptance by other departments will be increased.

Since the executives have asked for this analysis to be completed, it is assumed they are supportive. Expanding this to other departments beyond the small scope is important as the same techniques and technologies may expand into other departments beyond marketing, including product development and customer service and may include incorporating new ways to capture and report data. Part of what this project will be doing is creating a framework that can be adapted to predict different outcomes, such as amount willing to spend on a policy or likelihood of filing a claim.

The second part of the pilot plan is to review and identify the goals that this strategy will accomplish. The initial goal set by TIC is to identify which individuals are most likely to purchase the caravan insurance policy. The pilot will demonstrate that there is a set of characteristics of individuals who would be most likely to purchase a caravan insurance policy so that a list for marketing can be created. This project will also then be able to say, based on the factors presented, whether a specific individual would be likely to buy a caravan insurance policy or not.

The pilot plan will also involve the training of individuals on the technology and methods to implement the predictive analytics strategy. Initially this would include heads of departments or a head of training and development, with the potential to expand based on the needs and any increased scope of the analytics. Preliminary analytics may involve a group who are gathering, cleaning, and modeling data, as well as a group of testers who will ensure the models are valid.

The fourth point identified by Newton (2019) is to create a timeline for the implementation of the pilot program. With the initial completion of the predictive analysis completed within a certain time, testing whether this was an effective strategy or not, or if it is feasible to implement, would need to occur within a specified period. This will allow for judicious decision making of whether to proceed with the project or to abandon this pilot project and cut losses.

The initial data collection, data cleaning, and modeling may be quick for TIC. Ongoing modifications to the model and testing may take longer. The expectation should be put forth for how long to ensure the project is worth the time that is invested into it. This will also allow time to show if implemented as a regular part of the operation at TIC, how long similar analyses should be anticipated to take.

The final part of initiating a pilot program is to identify a formal project manager for this endeavor. With the resources invested in this program to ensure it is effectively run in a way that is well-managed and reproducible, with oversight by appropriate individuals to ensure valid outcomes of the project. The project manager would oversee the implementation according to the documented methods so as to ensure results are reproducible.

Once the structure of this data analysis project is shown to be successful there will be a resulting framework for further expansion within TIC. The team of individuals who have worked on the pilot will have been working as though this were not a pilot program, but an established program, with the opportunity to provide feedback along the way for streamlined processes.

**Model**

**Development**

The random forest model was created for the dataset provided by TIC using R and Rattle. The raw data was visually examined and there were no missing values and no values that appeared anomalous, such as negative numbers or exceedingly small or large numbers. There were 348 caravan policies of the 5822 policies, representing approximately 5.98% of the policyholders. The data was loaded into R and assigned to variable name “coil”. Headers were also added assigning the header names to colnames(coil) .

A preliminary random forest model was run with default settings of 500 decision trees and the error rate was generated. In Figure 1, the curve for value of 1 flattens at about 100. For the value of 0 it flattens at about 30. The ideal number of trees would be greater than 100, but can be left at the default of 500 as a greater number of trees is not detrimental to the model.

**Figure 1**

***Error rate for preliminary random forest of dataset “coil”.***Application

Description automatically generated with medium confidence

*Note:* the curve flattens at around 100 trees for values of 1.

The preliminary model had an error rate of 98.1 for positives and 0.6 for negatives, as seen in Figure 2. The AUC was 0.6831. Identifying where modifications to this model could be made, it was evaluated for the most important variables. The Importance function in Rattle produced a list of the variables that led to the greatest decrease in Accuracy and Gini. In Figure XX the graph shows the comparison of variables with those who have decreased accuracy the most at top and those least on the bottom. The list of variables on the y-axis is in inverse order for the variables identified as most important (Martinez-Taboada & Redondo, 2020). All variables that had a negative impact on accuracy and decreased the Gini score to a greater degree, noted in Figure 3, were removed and the model recalculated.

**Figure 2**

***Error matrix for random forest of dataset “coil”.***

***Graphical user interface, text, application, email

Description automatically generated***

**Figure 3**

***Visualization of mean decrease accuracy and mean decrease Gini for all variables.***

Chart, bar chart

Description automatically generated

**Figure 4**

***Lowest on lists of variables with negative mean decrease of accuracy and the corresponding mean decrease of Gini of the random forest model of dataset “coil”.***

Table

Description automatically generated with medium confidenceA picture containing text, receipt

Description automatically generated

Pruning the tree involved the removal of AMOTSCO – number of motorcycle/scooter policies, AGEZONG – number of family accidents insurance policies, PAANHANG – contribution trailer policies, ABYSTAND – number of social security insurance policies, MGODRK – living in zip code that is predominantly Roman Catholic, PMOTSCO – contribution motorcycle/scooter policies, AAANHANG - number of trailer policies, PZEILPL – contribution surf board policies, and ALEVEN – number of life insurance policies. These variables were then compared to each other to see if there are any patterns between them. It is noted that motorcycle/scooter policy data both decreased accuracy of the model. No other variables in this dataset were related to motorcycles or scooters.

The model of this initial pruning resulted in an improvement in error rate for positive values to 96.3. The error rate increased slightly to 0.9 for the negative values. The AUC increased significantly to 0.7232. With this random forest, importance of variables was again identified, with several having negative impacts on accuracy: AFIETS – number of bicycle policies, PLEVEN – contribution to life insurance policies, PWAPART – contribution to private insurance, PWALAND – contribution to agriculture insurance, PINBOED – contribution to property insurance, PBYSTAND – contribution to social security insurance, and ABESAUT- number of delivery van policies. Commonalities between these variables as well as the ones previous culled were not obvious.

These seven variables were removed from the model in this level of pruning. The resulting random forest model decreased in accuracy to an error rate of 98.1% for positive values. The negative values remained the same. The AUC also decreased to 0.70. Repeating this process only increased the error rates and decreased the AUC.

Continuing to prune the random forest by removing variables would not be ideal. Random forests typically do not need to be pruned in this way as they tend to not overfit the data when untouched (Krueger, 2022). Instead, tuning the hyperparameters, “the values that have to be evaluated with respect to their OOB performance” (Probst et al., 2019) will yield slightly improved AUC and accuracy. The authors continue by saying that the Gini score is a more helpful measure, along with the AUC, in building a more accurate model and encourage the use of the variable importance feature in Rattle, but that random forests rarely require significant tuning.

Other changes to the hyperparameters were completed with no impact on the overall functioning of the model. The partitions for training/validation/test data were also not effective. Changing the number of trees did not have an impact on the model. The final random forest model was the initial random forest model as other modifications for pruning and tuning had little, if any, positive impact on the model’s accuracy and validity.

**Performance**

The random forest model was scored using the provided eval data set in Rattle, using the score function in the evaluation tab, as seen in Figure 5. The txt file was loaded as a table, headers added, then converted to CSV using the R function write.csv().

**Figure 5**

***Rattle setup for scoring with “ticeval” data set.***

Graphical user interface, text, application, email

Description automatically generated

The error matrix produced by the eval data is seen in Figure 6 and the corresponding calculations in Figure 7. The random forest model predicted 1.05% of the values to be positive, while the actual positive rate was 5.95%. The model underestimated the positives but had a false positive rate of 0.0080. The overall accuracy of the model was 93.60%. While the model was conservative with assigning a positive value to rows of data, the overall accuracy of the model was high.

**Figure 6**

***Confusion Matrix of predicted values by random forest compared to actual values.***

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Total: 4000 | |  |  |  |
| Actual |  | Predicted | |
|  | **0** | **1** |
| **0** | 3732 | 30 |
| **1** | 226 | 12 |

**Figure 7**

***Confusion matrix calculations***

***Table

Description automatically generated with medium confidence***

*Note:* created with *Confusion Matrix - Online Calculator*. (n.d.). Retrieved October 16, 2022, from https://onlineconfusionmatrix.com/

**Conclusion**

Predictive analytics is a powerful tool to reduce the work and waste for a company like TIC. Tasked with identifying the individuals who are more likely to purchase caravan policies, a random forest model was created. The random forest model included all available data points, including demographic information and previous purchases. The outcome is a model that performed significantly better than random chance at identifying potential caravan policy purchasers.

This model would now be able to assess new data points and predict whether the corresponding individuals would benefit from a caravan policy. For TIC, this means marketing can be more specific and target only those individuals who are more highly likely to purchase the product. This will reduce wasteful spending on marketing campaigns that are not productive. As only 5-6% of the population of people are likely to purchase this product, instead of spending the marketing budget and time on the 94-95% that are not interested, the focus on the right individuals will be more cost effective as well as a better use of employee time and resources.

The models can be reworked to focus on different policy types, or identifying commonalities between certain policy holders. The updates and new data can also be run in real-time with exporting the model to other formats, such as those that will connect to data entered in a customer database. Overall TIC would benefit financially and employee work can be more effectively focused using predictive analytics.

**References**

Altman, R. (2017, November 21). *Most Effective Forms of Advertising*. Small Business - Chron.com. Retrieved September 11, 2022, from https://smallbusiness.chron.com/effective-forms-advertising-60755.html

Benes, R. (2018, March 23). *Marketers Waste About One-Fourth of Their Budgets*. Insider Intelligence. Retrieved September 11, 2022, from https://www.insiderintelligence.com/content/marketers-waste-about-one-fourth-of-their-budgets

Bhatia, N. (2021, December 10). *What is Out of Bag (OOB) score in Random Forest? - Towards Data Science*. Medium. Retrieved October 16, 2022, from https://towardsdatascience.com/what-is-out-of-bag-oob-score-in-random-forest-a7fa23d710

Burns, E. (2022, March 30). *Data Cleaning in R Made Simple - Towards Data Science*. Medium. Retrieved October 9, 2022, from https://towardsdatascience.com/data-cleaning-in-r-made-simple-1b77303b0b17

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0. step-by-step data mining guide*. SPSS. Retrieved September 4, 2022, from https://docplayer.net/202628-Crisp-dm-1-0-step-by-step-data-mining-guide.html

*Confusion Matrix - Online Calculator*. (n.d.). Retrieved October 16, 2022, from https://onlineconfusionmatrix.com/

DataRobot. (2022, July 20). *Scoring Data - DataRobot AI Cloud Wiki*. DataRobot AI Cloud. Retrieved October 16, 2022, from https://www.datarobot.com/wiki/scoring/

Dwilson, S. D. (2016, October 26). *What Demographics Respond the Most to Direct Mail?* Small Business - Chron.com. Retrieved September 11, 2022, from https://smallbusiness.chron.com/demographics-respond-direct-mail-66161.html

Elkan, C. (n.d.). *Magical Thinking in Data Mining: Lessons From CoIL Challenge 2000*. UCSD.edu. Retrieved October 16, 2022, from https://cseweb.ucsd.edu/~elkan/kddcoil.pdf

Gentleman, R., & Temple Lang, D. (2007, March). Statistical Analyses and Reproducible Research. *Journal of Computational and Graphical Statistics*, *16*(1), 1–23. https://doi.org/10.1198/106186007x178663

Harris, M. (2020, January 24). *GSA CDO: Data Outcome Feedback Loops Can Improve Predictive Analytics*. Government CIO Media. Retrieved October 16, 2022, from https://governmentciomedia.com/gsa-cdo-data-outcome-feedback-loops-can-improve-predictive-analytics

IBM. (2021, January 26). *Random Forest*. Retrieved October 16, 2022, from https://www.ibm.com/cloud/learn/random-forest

insightsoftware. (2022, July 13). *Top 5 Predictive Analytics Models and Algorithms*. Insightsoftware. Retrieved September 25, 2022, from https://insightsoftware.com/blog/top-5-predictive-analytics-models-and-algorithms/

Krueger, E. (2022, January 6). *Build Better Decision Trees with Pruning - Towards Data Science*. Medium. Retrieved October 23, 2022, from https://towardsdatascience.com/build-better-decision-trees-with-pruning-8f467e73b107

Larose, D. T., & Larose, C. D. (2015). *Data Mining and Predictive Analytics (Wiley Series on Methods and Applications in Data Mining)* (2nd ed.). Wiley.

Lee, C. S. (2016, October 3). *Logistic Regression versus Decision Trees*. The Official Blog of BigML.com. Retrieved September 25, 2022, from https://blog.bigml.com/2016/09/28/logistic-regression-versus-decision-trees/

Maraz, M. (2018, April 2). *Building a Predictive Lead Scoring System for Talent Agencies (part 1)*. Untitled Research. Retrieved October 16, 2022, from https://untitled-research.com/blog/building-a-predictive-lead-scoring-platform-for-talent-agencies

Marketing Evolution. (2022). *How to Use Predictive Analytics in Data-Driven Marketing*. Retrieved September 11, 2022, from https://www.marketingevolution.com/knowledge-center/the-role-of-predictive-analytics-in-data-driven-marketing

Martinez-Taboada, F., & Redondo, J. I. (2020, January 4). *Variable importance plot (mean decrease accuracy and mean decrease Gini).* Figshare. Retrieved October 23, 2022, from https://plos.figshare.com/articles/figure/Variable\_importance\_plot\_mean\_decrease\_accuracy\_and\_mean\_decrease\_Gini\_/12060105/1

Mitchell, M. (2021, December 11). *Selecting the Correct Predictive Modeling Technique*. Medium. Retrieved September 25, 2022, from https://towardsdatascience.com/selecting-the-correct-predictive-modeling-technique-ba459c370d59#

Newton, M. (2019, April 10). *Predictive analytics pilots that work*. Smart Water Magazine. Retrieved October 9, 2022, from https://smartwatermagazine.com/blogs/matt-newton/predictive-analytics-pilots-work

Probst, P., Wright, M. N., & Boulesteix, A. (2019, January 28). Hyperparameters and tuning strategies for random forest. *WIREs Data Mining and Knowledge Discovery*, *9*(3). https://doi.org/10.1002/widm.1301

Sentient Machine Research. (n.d.). *CoIL Challenge 2000 Report*. Retrieved October 12, 2022, from https://liacs.leidenuniv.nl/%7Eputtenpwhvander/library/cc2000/data.html

Sharma, G. (2022, August 2). *5 Classification Algorithms you should know – introductory guide!* Analytics Vidhya. Retrieved September 25, 2022, from https://www.analyticsvidhya.com/blog/2021/05/5-classification-algorithms-you-should-know-introductory-guide/

The Data Company. (2020, March 24). *Benefits and examples of Predictive Analytics*. Retrieved September 11, 2022, from https://thedatacompany.com/benefits-and-examples-of-predictive-analytics/

Williams, G. (2010). *DATA MINING Desktop Survival Guide*. togaware.com. Retrieved October 16, 2022, from https://datamining.togaware.com/survivor/Scoring.html

Williams, G. (2011). *Data Mining with Rattle and R: The Art of Excavating Data for Knowledge Discovery (Use R!)* (2011th ed.). Springer.