**Predictive Modeling Final Project**

Cameron Gallien

DSE6111OM\_SU2023S5\_Predictive Modeling

Merrimack University

Prof. Fotios Kokkotos

August 25, 2023

**Executive Summary**

This report presents the application of various algorithms and techniques to address real-world problems using customer survey data collected from a telecommunications company. The main objectives of this project were to predict customer tenure length, equipment rental decisions, and monthly spending amounts for both current and prospective customers of the company. The goal of this report was to evaluate and compare the performance of multiple models, and provide recommendations based on the outcomes.

For the quantitative response objective, predictive models including linear regression, ridge regression, lasso regression, partial least squares regression (PLS), regression trees, bagging, random forests, and boosting were all explored. Subset selection methods were used to determine relevant predictor variables, and model evaluations were based on metrics like mean-squared error (MSE). Among the models, the random forest with 15 variables showed the lowest test MSE of 45.71.

The qualitative response objective aimed to predict whether or not customers would decide to rent telecommunications equipment from the company. Logistic regression, linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), k-nearest neighbors (KNN), classification trees, bagging, and random forests were employed. The random forest model with 2 predictors ended up achieving the best results out of all the models considered, with an accuracy of 84.32% in predicting customer equipment rentals.

Finally, a principal components regression (PCR) model was developed to forecast monthly spending amounts for each customer. The model utilized standardized predictors and determined the optimal number of principal components through cross-validation. However, with an average test error of 19.242, further refinement is recommended to enhance the practical utility of this model.

**Data & Approach**

The data that was selected for this project was customer survey data collected by a telecommunications company. It contained 5000 observations and 60 variables. To begin, I selected specific variables that would need to be eliminated for a variety of reasons. These variables either did not contain useful information, were lacking context to understand what they represented, or would interfere with the goals of the project. There were also variables that needed to be removed specifically for each problem that was proposed. For more information on the variable selection process and the reasons for excluding specific variables, please refer to Appendix B.

In addition, many variables were simply encoded with the values of “1” or “0” for “yes” and “no” respectively. These variables were converted to factors so that they would not be viewed as quantitative variables. Information on these variables is also included in Appendix B.

Data was then separated into training and test sets, with 3750 observations (75%) assigned to the training set, and 1250 observations (25%) assigned to the test set. Each model was fitted on the training set, before evaluating its performance on the test set to measure how well it generalizes to new, unseen examples.

My goal on this project was to create as accurate a model as possible to answer three different analytic objectives using this data. The three problems considered were:

* Predict tenure length of a customer or prospective customer using various quantitative response models.
* Predict whether a customer or prospective customer would rent equipment from the company using various qualitative response models.
* Predict the total amount per month that a customer or potential customer would spend with the company using a Principal Components Regression model.

My reasoning for choosing these objectives was that this information would be extremely helpful for both targeted customer acquisition and retention efforts, as each objective could help the company obtain as much revenue as possible. For example, specific customers who are not expected to have long tenures with the company, or are expected to spend more money per month than the average customer could be targeted for loyalty programs to stay with the company for as long as possible.

For the quantitative response objective, the models used were linear regression (utilizing multiple different subset selection methods), ridge regression, lasso regression, partial least squares regression (PLS), regression trees, bagging, random forests, and boosting. Models used for the qualitative response objective were logistic regression, linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), k-nearest neighbors (KNN), classification trees, bagging, and random forests. The final objective was solved by using a principal components regression model (PCR) as noted above.

**Findings & Model Evaluation**

Quantitative Response: Customer Tenure

I sought out to see if there was a way to predict customer tenure based on the available variables in the customer survey data frame. The response variable for this objective is Phone Company Tenure. After removing the predictor variables mentioned in the Data & Approach section, my first step was to perform multiple different subset selection techniques to determine the ideal variables and number of variables to be used in my linear regression models. The subset selection methods used were best subset selection, forward stepwise selection, backward stepwise selection, and in addition, validation-set and cross-validation approaches were also utilized.

I began by plotting the residual sum of squares (RSS), adjusted R2, C­­p, and Bayesian Information Criterion (BIC) metrics that were produced from the best, forward, and backward stepwise selection methods (these plots can be found in Appendix A, Figures 1-3). All three of these methods produced almost identical results, with the RSS metric decreasing, and the adjusted R2­ metric increasing as more variables were included. In addition, the BIC metrics for all three subset selection methods were at their minimum point when 12 variables were included. According to James et al. (2021), the BIC metric is generally the lowest when a model has a low test error, making it the typical choice for utilization. Because of this, I decided to move forward with this model for my linear regression models.

In addition to all methods choosing a 12 variable model, all three methods determined that the same 12 variables should be utilized. These variables were age, education years, marital status, tenure with their credit card, amount spent on voice services and data services last month (both in dollars), whether the customer owned a fax machine, whether they were a news subscriber, and whether the customer was utilizing calling card, multiline, and call forwarding services. In addition, the validation-set approach for subset selection produced identical results, choosing the same 12 variables to be used in linear regression. Additionally, these methods showed that there were 15 variables that did not appear in any of the ideal model sizes up to size 26. Because of this, these variables were removed from the remainder of the quantitative analysis models due to their lack of influence to offset the long amount of time that the computer was taking to process them.

The training set of the 12 selected variables was then fitted to a linear regression model, which was evaluated using the test set. This linear model produced a mean-squared error (MSE) of 71.349. This showed promising results as a potential model to be used to answer the objective.

The cross-validation subset selection method was used as well, with this method determining that all 26 remaining variables should be utilized. A linear regression model was then fitted to these 26 variables, producing a test MSE of 71.678, slightly higher than that of the previous model.

The next model utilized was a ridge-regression model. I used cross-validation to determine the ideal value for λ. 10-fold cross-validation produced a value of 2.062 for this metric, which was then utilized in the model. The ridge-regression model produced a test MSE of 77.471, which was significantly higher than that of the previous two models.

A lasso regression model was then fitted to the data, using the same cross-validation method to determine the ideal value of λ. Cross-validation for the lasso model produced a much smaller value for λ, .016. Just as was done previously, this value was incorporated into the lasso model which produced a test MSE of 71.715. This value was smaller than that of the ridge regression model, but still higher than that of the first two linear models.

Next was a PLS model. This model also utilized cross-validation to determine which value of *M* (the number of principal components used) would provide the lowest test error. This value was determined to be 9, as this correlated to the lowest root mean squared error of all values of *M* (8.649). This *M* value was inputted into the model, which produced a test MSE of 71.65, which was pretty much in line with the test errors that had been reported thus far.

The next set of models used were tree-based, the first being a regression tree. A regression tree was fitted on the training data (Appendix A, Figure 4) before being tested on the test data. This tree showed a test MSE of 89.347, by far the highest of any of the model so far. I then used cross-validation to see if a pruned tree would lead to improved results, however upon plotting the error rate as a function of tree size, it showed that the un-pruned tree correlated with the least amount of cross-validation errors (Appendix A, Figure 5).

The next tree-based models used were random forests. The first one used the bagging method which is a random forest with all 26 predictors used. This resulted in a test MSE of 47.301, a very substantial improvement over all the models used so far. Additional random forests were also produced to try to reduce the test MSE even further. First, a random forest was built using the default settings for the amount of variables used (p/3). This produced a test MSE of 46.532, which showed even more improvement. I then began experimenting with different numbers of variables and found that the lowest test MSE I could obtain was by utilizing 15 variables. This resulted in a test MSE of 45.71, the lowest thus far.

The last method used for this objective was boosting. After experimentation, I found that the lowest MSE obtainable was with 100 trees. This model produced a test MSE of 47.304, still low, but not as low as the random forest with 15 variables, which produced the lowest test MSE at 45.71.

Qualitative Response: Equipment Rental

The next objective to tackle was to predict the Equipment Rental variable (a factor, “1” for “yes” and “0” for “no). I started by once again removing the variables mentioned in the Data & Approach section, as well as the variable for the amount of dollars spent on equipment over a customer’s tenure. This would evidently introduce bias into the model’s predictions seeing as only those who have rented equipment would have a value attributed to this variable. For each model used, predicted probabilities were computed as to whether each test set customer would rent equipment. If the probability of a customer renting equipment was above 50% it was labeled as “yes”, if it was below 50% it was labeled as “no.” These were then compared to the actual value for each test customer to determine the test accuracy of each model.

The first classification model that was fitted to the data was a logistic regression model. Once the model was fitted to the training data, predictions were collected and compared to the test set. The logistic regression model had an 83.44% accuracy rate, a promising result for my first classification model.

The next two models were LDA and QDA models. These models also showed very promising results, with the LDA model producing an accuracy of 83.12%, and the QDA model producing an accuracy of 81.12%. These were both lower than that of the logistic regression model, but still seemed to be working well at predicting equipment rentals.

Next was the KNN model. This model was fitted to the training data as were the previous, and after some experimenting with *K*-values I was able to get the accuracy of this model up to 75.76% by using a value of *K*=15. However, since these results were less than optimal considering the previous models that were attempted, I then standardized the variables so that they were all on the same scale. This would prevent variables with higher amounts from outweighing the others. To my surprise, this revealed less than stellar results, as I was only able to increase the accuracy to 76.72% with a *K*-value of 25.

After the KNN models were fitted, I then fitted a classification tree (Appendix A, Figure 6). This tree resulted in an accuracy of 80.4%. I then used cross-validation to see if a pruned tree would provide better results, but once plotted the cross validation showed that an un-pruned tree had the lowest amount of cross validation errors (Appendix A, Figure 7).

The final step of this objective was to fit the random forest models. The first one fitted was done using the bagging method and produced an accuracy of 81.6% on the test data. Next, a random forest was built using the default settings for the number of variables used (√p). This produced an accuracy of 83%, which gave me optimism that with some tuning I could increase this metric. This ended up becoming reality when, after some experimentation, I was able to find that by using only 2 predictors the accuracy was raised to 84.32%. This would ultimately be the highest accuracy achieved for this objective and is recommended for use in solving this problem.

Principal Components Regression: Amount Spent per Month

The final objective remaining was to try to predict the amount a current or prospective customer would spend on a monthly basis with the company using a PCR model. I started by once again removing the variables mentioned in the Data & Approach section and converting the applicable columns into factors. Since there were three different variables that encompassed how much a customer spent in the last month (separated by voice, equipment, and data charges) I created a new variable to be used as the response variable that consisted of the sum of these three columns. The three separate columns were then were removed so that they would all be encompassed within one variable. I also removed the equipment rental variable, as this would introduce bias, and would already be represented in the total amount spent variable.

In fitting the model, the remaining 38 predictors were standardized to avoid the scale of each predictor influencing the final results. Then, cross-validation was utilized to try to find the value of *M* (the number of principal components used) that would result in the lowest test error rate. Once the cross-validation scores were plotted (Appendix A, Figure 8) and analyzed, it was apparent that the lowest test error occurred when all 38 predictors were used in the model. This *M*-value was inputted into the PCR model, which resulted in an average test error of 19.242. However, since the median amount spent per month through the telecommunications company is only $24.55, it is advised that further work be done to find a more fitting model as the results may not be ideal in practice.

**Validity & Reliability Assessment**

The analysis-driven recommendations that were presented in this report provide a lot of valuable insights into the behavior of those customers that completed the survey. However, the validity and reliability of these recommendations depends on several factors.

The first important factor is the accuracy of the customer survey data. Since this information was filled out by individual customers, there is some concern as to the accuracy of what was reported. Going forward, data collection of future customers will not be as easy. It may be helpful to have customers fill out a survey when signing up with the company to collect some of this information from newer customers.

In addition, these models were trained on a specific dataset, and may perform differently on new, unseen data. The recommendations included in this report should be tested in real-world scenarios in order to properly assess their accuracy and effectiveness. Monitoring their performance and making adjustments based on the results of real-world testing is crucial to confirming and improving the reliability of these models.

In conclusion, the findings and recommendations discussed in this report offer some valuable insights into customer behavior and business decisions. That being said, the validity and reliability of these recommendations rely on many different factors such as those previously mentioned. Continuous evaluation, refinement, and consideration of real-world implications are very important to make sure that the models serve their intended purpose, and that implementation of these models serves its intended purpose.

**Appendix**

Appendix A: Additional Graphs

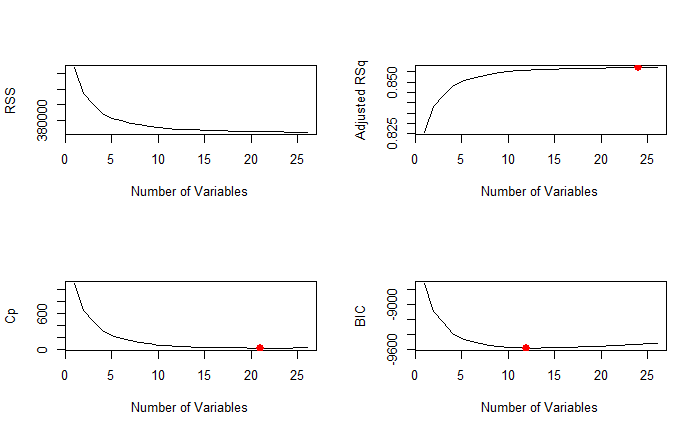


Figure A.1: Residual sum of squares (RSS), adjusted R2, C­­p, and Bayesian Information Criterion (BIC) plots for best subset selection.

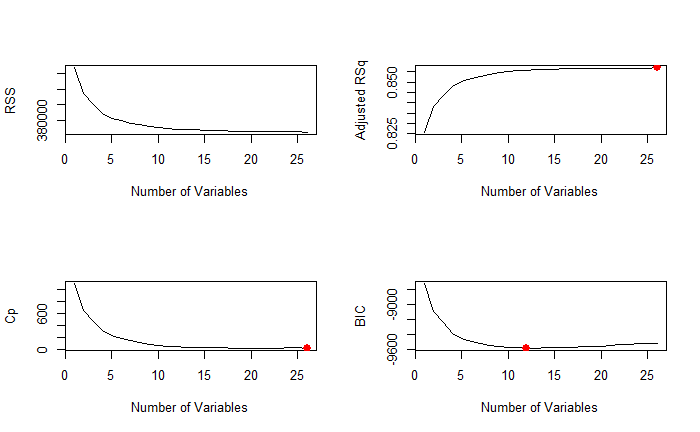


Figure A.2: Residual sum of squares (RSS), adjusted R2, C­­p, and Bayesian Information Criterion (BIC) plots for forwards stepwise selection.

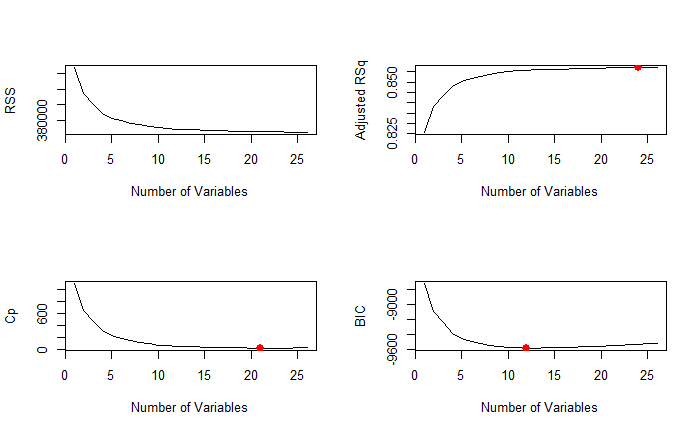


Figure A.3: Residual sum of squares (RSS), adjusted R2, C­­p, and Bayesian Information Criterion (BIC) plots for backwards stepwise selection.

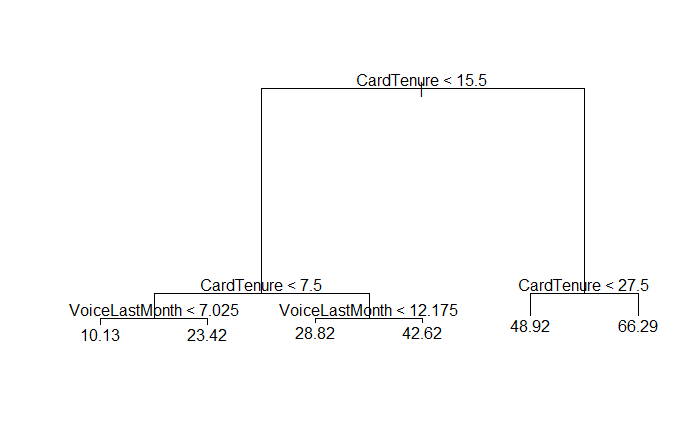


Figure A.4: Quantitative analysis regression tree.

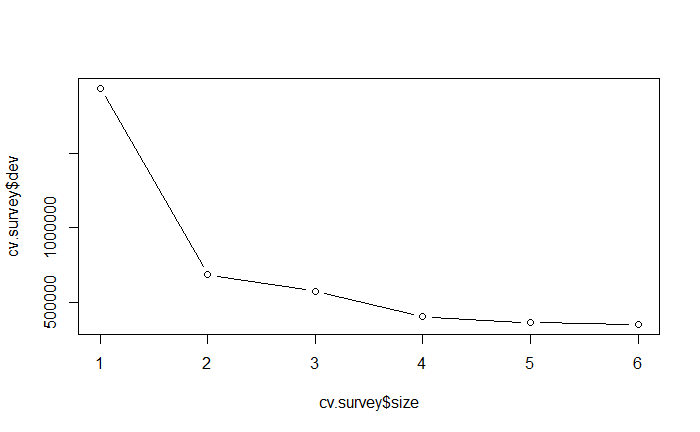


Figure A.5: Error rate as a product of tree size for the regression tree.

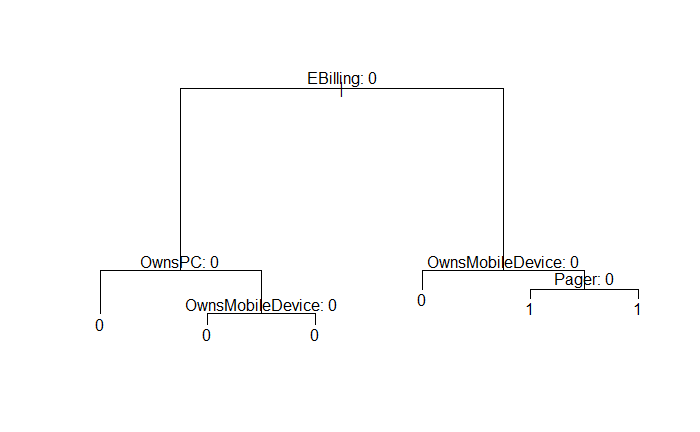


Figure A.6: Qualitative analysis classification tree.

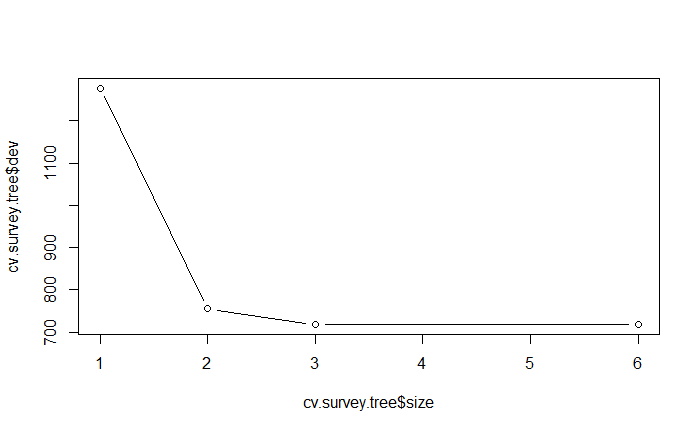


Figure A.7: Error rate as a product of tree size for the classification tree.

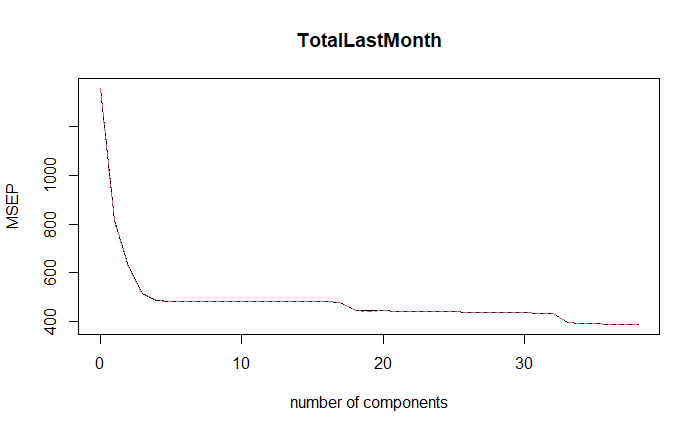


Figure A.8: Mean Squared Error of Prediction (MSEP) as a product of the number of components used in PCR model.

Appendix B: Variable Selection

The variables listed below were excluded from the models in this paper due to a variety of reasons.

* Only N/A values present: Telecommute
* Arbitrary information: Customer ID
* Only non-binary numerical values existed with no information to their context: Region, Job Category, Car Ownership, Car Brand, Town Size, Credit Card, Internet
* Variables already represented within other variables:
  + Number of Cats, Number of Dogs, Number of Birds (represented in Number of Pets)
  + Credit Debt, Credit Card Items Monthly, Other Debt (represented in Debt to Income Ratio)
* Would introduce bias in regards to all response variables: Voice Spent Over Tenure, Equipment Spent Over Tenure, Data Spent Over Tenure (all in dollars).

In addition, each objective required specific variables to be excluded.

* Quantitative Response: Customer Tenure
  + 15 variables were excluded due to their negligible impact on results: Retired, Household Income, Number of Pets, Homeowner, Cars Owned, Car Value, Commute Time, Political Party Member, Voting Status, Card Spend per Month, Voicemail, Caller ID, Call Waiting, Three Way Calling, EBilling.
* Qualitative Response: Equipment Rental
  + 1 variable was excluded to prevent the introduction of bias: Equipment Spent Last Month (if a customer did not rent equipment this would be 0)
* Principal Components Regression: Amount Spent per Month
  + 3 variables excluded because they were already represented in the new variable Total Spent Last Month: Voice Spent Last Month, Equipment Spent Last Month, Data Spent Last Month
  + 1 variable excluded to prevent the introduction of bias: Equipment Rental (those who rented equipment had non-zero amounts in this variable which was represented in Total Spent Last Month)

Lastly, the following variables were converted to factors as they contained only binary values correlating to “yes” and “no”: Gender, Union Member, Retired, Loan Default, Marital Status, Home Owner, Political Party Member, Voting Status, Active Lifestyle, Calling Card, Wireless Data, Multiline, VM, Pager, Caller ID, Call Waiting, Call Forward, Three-way Calling, EBilling, Owns PC, Owns Mobile Device, Owns Game System, Owns Fax, News Subscriber.

**References**

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An Introduction to Statistical Learning with Applications in R (Second Edition). Springer