Coursework -2,

MSCI-562

Author: 36523348

Introduction and Feature Selection:

This report builds on a previous exploratory analysis of features used to determine the customer churn results of whether customers have exited the bank or not. It focuses on feature selection followed by testing several classification models and selecting the best one to predict customer churn towards retaining customers in the future.

In the previous coursework, the variables, and their impact on distinguishing customer churn classes (Exited – Yes and No) was evaluated, along with the suitability of the features in being included for classification, using visualizations such as conditional bar and boxplots. Amongst all, the quantitative features with better discrimination power were "Age" and "Balance", and amongst the factor variables were "Complain", "Is Active Member", "Geography", "Number of Products" and "Gender".

An in-depth analysis revealed that the Balance feature was zero inflated and carried an anomaly, where some customer who had exited the bank still had a large balance in their accounts. However, this variable is maintained and will be analyzed for effects on classification. Additionally, "is Active Member" variable contained many discrepancies, some illogical, such as some members (15% of training dataset) who had exited the bank were deemed "active". Due to the unreliability of this variable, it is excluded from the analysis, as even good classification results cannot produce a reliable model from illogical variables. This leaves a total of 6 features to be used for classification.

Classification

1. k-Nearest Neighbours classifier

For this method, two important models will be discussed: (i) Model-1 consisting of the quantitative variables from the feature selected dataset (Age and Balance). (ii) Model-2 consisting of all the variables from the feature selected dataset.

Model-1: Model with Quantitative Variables for kNN-classification.

Rationale: This classifier is a distance-based classifier; hence it would make sense to consider the quantitative variables from selected features for classification. While it would impact the results negatively, foregoing categorical features, it would also provide important insights into the extent of classification and uncertainty involved, and the improvements needed therein without making the model overly complex incorporating all known variables from the start, as kNN is computationally expensive and suffers in high dimensionality. KNN does not make strong assumptions about the underlying data distribution, making it suitable for the zero-inflated variable available to us. Additionally, Age and Balance are not well separated in terms of resultant classes of customer churn, as shown in Figure 1 – A, It can handle non-linear and complex decision boundaries without requiring prior knowledge about the data, hence even the complicated and somewhat merged datapoints of Age and Balance can be handled non-linearly, making this a model that serves as a baseline model for classification tasks and can provide quick insights into the data.

Methodology: The training data was divided into a training set (80%) and a test set (20%) to evaluate the model's performance, before testing on unknown data (actual test set or data set 2). Various values of k (1 to 50) were tested to find the optimal setting for classification, since this is a supervised learning algorithm based on pair-wise similarities, hence the optimal k will influence the results radically. Due to the imbalanced classes of the resultant labels in the training set (ration of exited yes to no is 0.25, hence only 25% of the training data are customers who have exited, and 75% have not exited), we used the balanced accuracy measure, which is the average of sensitivity and specificity, to assess the model's performance. Normal accuracy measures can easily misguide the results, as assigning class "Not exited" to the entire dataset can easily produce a normal accuracy measure of 75%, but the balanced accuracy will be much lower hence more representative of the true results. Additionally, kNN does not require feature scaling or normalization, hence we can proceed with the classification directly.

Results and Discussion: The balanced accuracy, sensitivity and specify scores for different values of k, are presented in the accompanying figure. The classification results were consistent with our expectations, as the data points were not easy separable into resultant classes, visible in their scatter Figure 1 - A, the model exhibited low accuracy with a the most complex k value (k=1), achieving approximately 55% accuracy on the test set, accuracy only declined further on consecutive k values, shown in Figure 1 - B. Since random classification can provide an accuracy of 50%, this model is not worth more than a baseline. However, it is indicative of the fact that more features require incorporation for improved classification accuracy, and a simplistic model with two quantitative variables will not be sufficient.

Model-2: Mixed Variable model using Gower Distances for kNN classification.

Rationale: Previously it was concluded that the model required additional variables to provide useful classification. Hence, important categorical variables with classification potential identified in the previous coursework, were included in the model i.e "Complain", "Is Active Member", "Geography", "Number of Products" and "Gender". Since this is a mixed variable model, and kNN requires distance measures as datapoints, Gower's Distance was used to calculate the dissimilarity between variables based on the proportion of mismatches between categories. Incorporation of additional important variables is expected to improve the classification output and provide a usable model.

Method: The training dataset was used to calculate dissimilarities using Gowers and scaled to interpretable two dimensions. However, Gower's distance normalizes the dissimilarities from different variable types to ensure that each variable contributes equally to the overall distance. Additionally, multidimensional scaling of the dissimilarity matrix returns reflection distances approximately equal to dissimilarities. In such a case the separate scaling of test and train data can produce results in different scaled systems. To counter this issue, a merged dissimilarity matrix was created by combining the test and train data set, to obtain datapoints in similar scales, and then split into training and test data sets for kNN classification. Results of both; separate and combined dissimilarity matrices are described.

Results and Discussion:

The result of separate dissimilarity matrices of train and test datasets have separate scale ranges, as shown in Figure 1 – D and E. The test set has a y axis scaled from -0.2 to 0.4 while the test set has -0,4 to 0.2. A model trained on train set will not be appropriately applied to the test set, and the out-of-range scaling will affect the classification results. This is apparent in Figure 1 – F, where the classification results at any k value, do not go above approximately 46% accuracy. This led to the creation of a merged data dissimilarity matrix, which was separated into test and training data, and used in the kNN classifier. This technique scaled the data in the same range and allowed the correct application of the training model onto the test data. From the scatter plot of the training data Figure 1 - C, we can see a distinct separation of classes as separate clusters. This is indicative of a good data set for kNN as clear separation boundaries can be created leading to strong predictive power. The model was tested at different k values for the optimal balanced accuracy, results compiled in Figure 1 - F. The best accuracy results of 99% are obtained at k = 5.

For kNN, a mixed variable model offers optimal performance, although on a simple machine it is very computationally expensive, hence feature selection would be pivotal for this technique. Although with this technique we could not determine the optimal feature set, it is expected that regression will allow us to identify the variables contributing to resultant class differentiation significantly and allow the selection of an even smaller subset of features. Additionally, the inclusion of training and test sets for dissimilarity matrices does not seem like an optimal technique, as it ends up making the process computationally expensive for each new test set obtained, as each new test set would require the same iterative process, thus making it computationally infeasible. However, in terms of generalization, this technique would perform well on all new datasets with the similar features in focus.

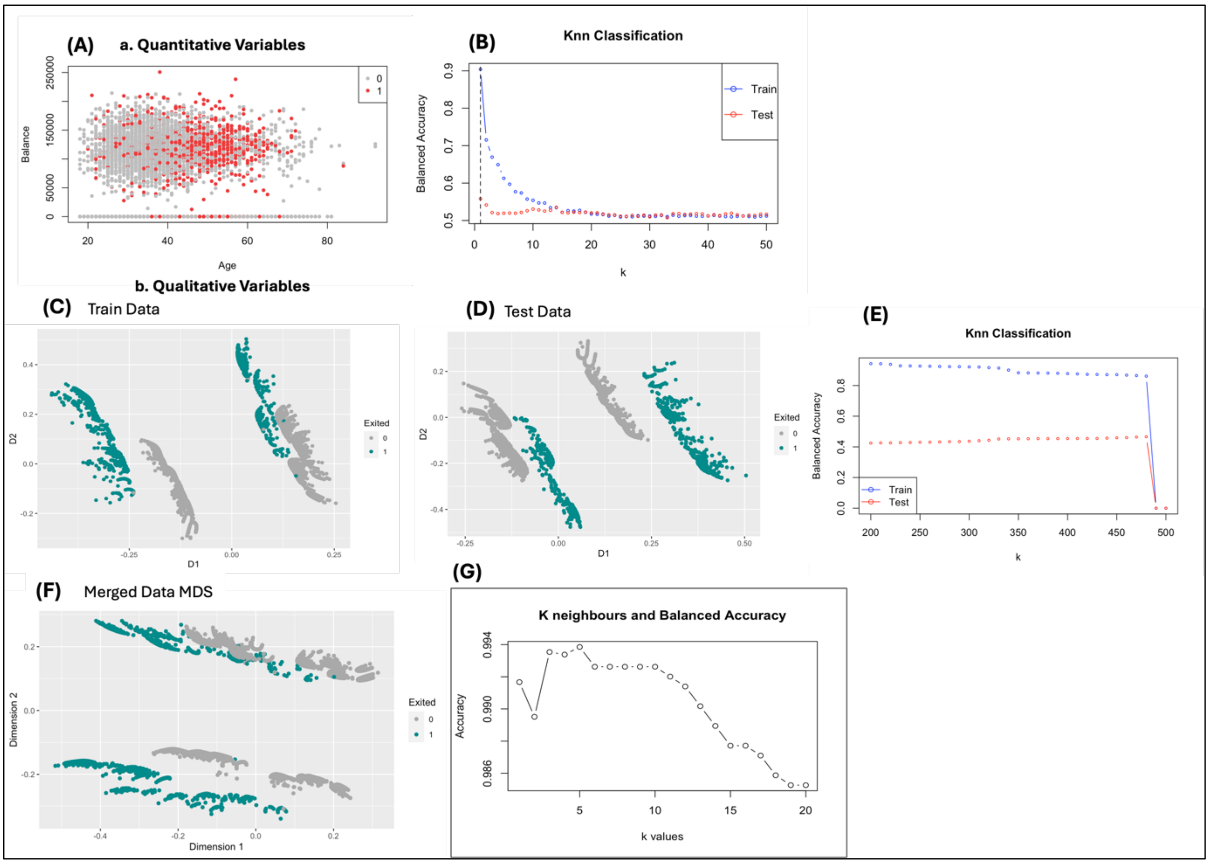


Figure 1 Results of kNN models (A) Scatter plot of variables "Age" and "Balance" (B) Balanced accuracy values of training set at varying k values of nearest neighbours (C-E) MDS scaling of variables using separate dissimilarity matrices and associated accuracy results after training at varying k values (F-G) MDS scaling of varibales using a merged dissimilarity matrix and associated accuracy results after training at varying k values.

2. Regression Classifier

For this method, two important models will be discussed: (i) Model-1 consisting of the 6 features in consideration (ii) Model-2 consisting of a subset of variables with strong classification ability.

Model-1: Complete Variables Regression Model (6 variables)

Rationale: Regression-based classification models can accommodate a variety of variable types, including categorical, and numerical variables and provide easily interpretable coefficients that quantify the relationship between predictor variables and the outcome. This makes this technique a good choice for classification models and allows for a clear understanding of how each variable contributes to the classification decision, leading to subset selection for low complexity but well performing models. Additionally, they allow for regularization through penalization, but that won't be necessary in our case as we have already narrowed down the feature set to 6 variables. Limitations include collinearity and complex non-linear relationships in data. The features selected do not correlate (tested through Pearson's for quantitative and Cramer's V for categorical variables) and complex non-linearity will be tested through evaluation of results, since it is difficult to visualize for categorical variables.

Method: General linear model of the binomial family was used, as the dependent variables "Exited" has a binomial distribution of the type 'yes' or 'no'. Two regression methods were particularly useful, one was complete regression with all 6 variables, and another was subset selection. Confusion matrix of results was compiled to calculate sensitivity, specificity and balanced accuracy of model results on training and test datasets.

Results and Discussion: The complete regression model indicated that all the variables were not significant to the model, in fact just the 'Age', 'Number of Products' and 'Complain' variables were of significance, as indicated by the p-values of coefficients (less than 0.001), implying that these variables contribute significantly in influencing the dependent variable "Exited". The complete model performed well in creating a classification model from the training data, that can be applied well on the test data. It gave a balanced accuracy of 99% on training data, as well as on the test data, at threshold 0.5.

Model-2: Subset Regression Models – 3 Variables best subset to 1 variable.

Rationale: Now that a model performs well above the required limit of identifying 80% of customer churn, the next improvement would be to simplify the model whilst maintaining classification accuracy. Cross-validation was used in conjunction with subset selection to assess the performance of different subsets of variables. This is done to create a simpler and more interpretable models by identifying a subset of predictor variables that have the most significant impact on the response variable. By reducing complexity, this model aims to reduce overfitting and increase generalizability. It may even improve model performance, but since we have a balanced accuracy of 99%, improvement is not the primary goal, rather it is the maintenance of this error metric. Additionally, feature elimination may improve the realistic applicability and interpretability of the model as well, as some features are problematic (such as zero inflation in balance).

Method: The 'bestglm' function was used for selecting the best subset of predictor variables for a regression model (Model 2A). It employed "subset selection" to systematically evaluate different combinations of predictors and identify the subset that produces the best-fitting model, through Akaike Information Criterion metric. AIC is preferred over BIC in this case as it is less stringent in penalizing complexity, and we aim to create a simpler subset whilst strictly maintaining accuracy. Cross validation is included in model selection with a common standard of parameters; 10-fold and 2 repetitions. The model is trained and applied to test data. It is evaluated by forming a confusion table of both, and determining the sensitivity, specificity and balanced accuracy of both training and test sets. Secondly, a more stringent penalty measure of LASSO is utilized to see if we can narrow down the model further whilst maintaining our classification evaluation goals (Model 2B).

Results and Discussion: The best subset model (Model 2A) indicated that only three features are essential to our classification objective, (i) Age of customers (ii) Number of Products bought with the credit card and (iii) Complaint status. This subset can mimic the training and test set classification accuracy of the previous more complex model and gives similar well separated prediction probabilities on the test set, hence a similar ROC curve and a balanced accuracy of 99% on the training and test data (Similar results as Model-1, hence not compiled again in figures). This simpler model tends to mitigate overfitting by reducing the feature complexity, while cross-validation ensures that the selected subset generalizes well to new data, improving or maintaining the model's performance in this case.

Amongst the essential features, the intercept includes the combined impact of Customers from France, customers who've bought 1 product and have not filed any complaints, collectively decrease the log odds of exiting the bank by -9.9. While filling a complaint increase the log odds of exiting the bank by 13 units, Age by 0.072 units and buying 2 products decreases the log odds by -2.28. This concludes that the strongest impact is caused by whether the customer has filed a complaint or not, and indeed the "Complaint" feature is a promising choice for classification model, as it shows an almost complete distinction in separating resultant classes of customer churn (Conditional Bar Plot). These coefficients indicate that all the variables utilized are not equally significant and some can be dropped whilst maintaining approximate accuracy goals of predicting 80% customer churn. For this, we proceeded to use the more stringent LASSO method of penalization and subset selection.

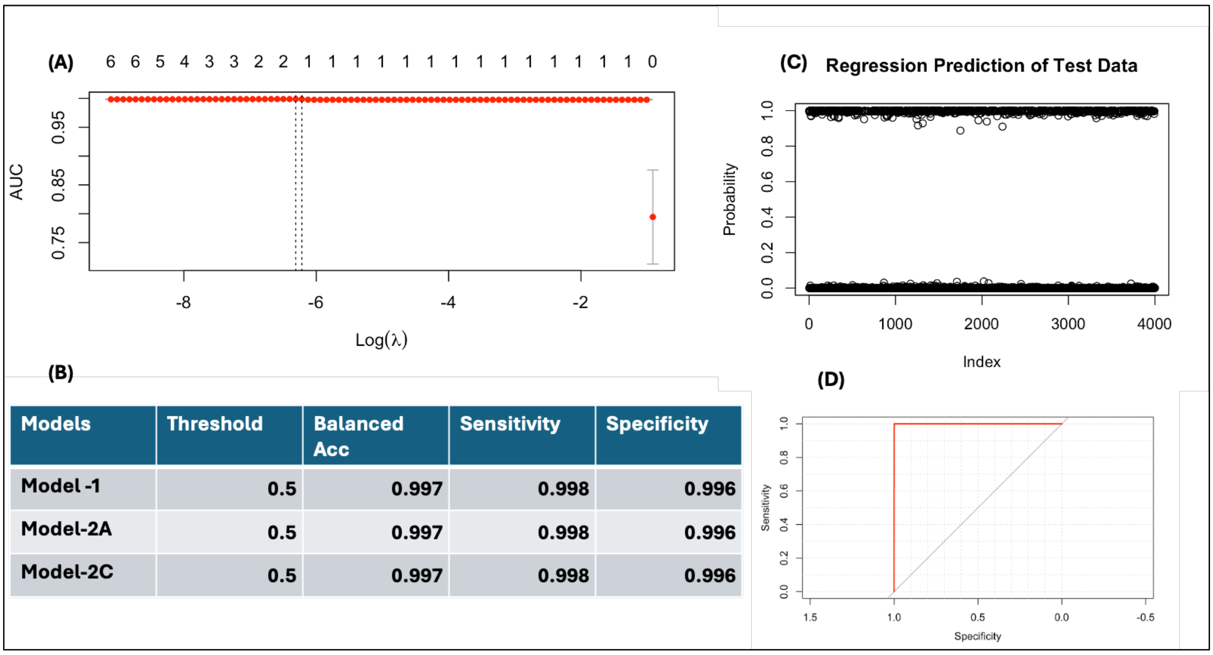
The LASSO model (model 2B) was cross validated for the optimal lambda penalization. The plot Figure 2 – A, exhibits result of cross validation. We can see that very stringent lambdas still effectively give 99% ROC on the training data set, event as the coefficients reduce to 1, but dropping significantly at log of lamda 0.89 having 0 coefficients. 0 coefficients in this case includes the intercept and one of the classes of Complain variable. This plot is evident of only one variable contributing significantly to the ROC evaluation, indicating that a one variable model can essentially produce the desired classification accuracy. The dashed line lambda model from plot, which points to a 2-variable model of "Age" and "Complain" features. We instead proceed to check the 1 variable model and recreate the regression model with "Complain" feature only (Model-2C).

As evident from the LASSO lamdas plot, the on variable model proved to equate the balanced classification accuracy of the most complex model up to three decimals, easy surpassing the required criteria of predicting 80% customer churn. This model produced well separated prediction probabilities and gave a balanced accuracy of 99% at 0.5 threshold of class separation. It can be stated as:

$P(\text{Exited} = 1 | \text{Age} = x, \text{Balance} = y \ldots, \text{Beta}_i = z) > 0.5$ and

$1 + \exp(9.98 - \Sigma\beta_i x_i) < 2$

However, the threshold was also varied, and results showed that the predicted class probabilities on test data were very well separated, as seen plotted in Figure 2 – C. It indicated that a class separation threshold of 0.1 to 0.8 would give equally good results, hence giving an extreme ROC curve increasing from 0 to 0.9 in sensitivity while maintaining specificity of 1, coherent with the threshold-based sensitivity and specificity results obtained manually Figure 2 – D. In this model the customers who have complained increase the log odds of exiting the bank with 12.4 units. This model has a good generalizability, simplest complexity, and realistic applicability whilst giving a similar prediction accuracy, as shown in Figure 2 – B. Hence amongst the regression models, Model 2C is recommended.



Models	Threshold	Balanced Acc	Sensitivity	Specificity
Model -1	0.5	0.997	0.998	0.996
Model-2A	0.5	0.997	0.998	0.996
Model-2C	0.5	0.997	0.998	0.996

Figure 2 Regression Classification Results on Test Data Set (A) The cross-validatioin plot of LASSO in Model 2C indicating the coefficients and area under the curve at each lambda (B) The evaluation metrics of predictions at a 0.5 threshold of class distribution of each model (C) Scatter of Probability predictions of test data belonging to Customer churn "Exited -yes" (D) ROC curve of train data predictions.

3. Decision Tree Classifier

Rationale: Decision trees are easily interpretable, can handle both numerical and categorical data without requiring extensive preprocessing and they can handle missing values and outliers naturally without the need for imputation or transformation. Additionally, they provide relative importance of each feature in the classification process and can capture nonlinear relationships thus accommodating complex decision boundaries. They are scalable, efficient and can be parallelized easily, enabling distributed computing on big data platforms. This makes it an intuitive choice to apply to our classification problem in question.

Method: A simple decision tree was created using the training data and inbuild R functions. Since the model created pure decision nodes classifying the training set with 1 branch, no pruning or parameter adjustment to limit complexity was required.

Results and Discussion: The decision tree created a simple one branch tree classifying the training data effectively, eliminating all other features while providing a balanced accuracy of 99% on training data at 0.5 threshold of class division. It was fast, efficient, simple and the results are easy to understand, as shown in Figure 3. Using the "Complain" feature as branching criterion, training data can be easily split into 20% customers who will exit if they have complained, while 80% of the customers will not exit who have not complained. The predicted classes are well separated as seen in the ROC curve besides it and explained in detail previously.

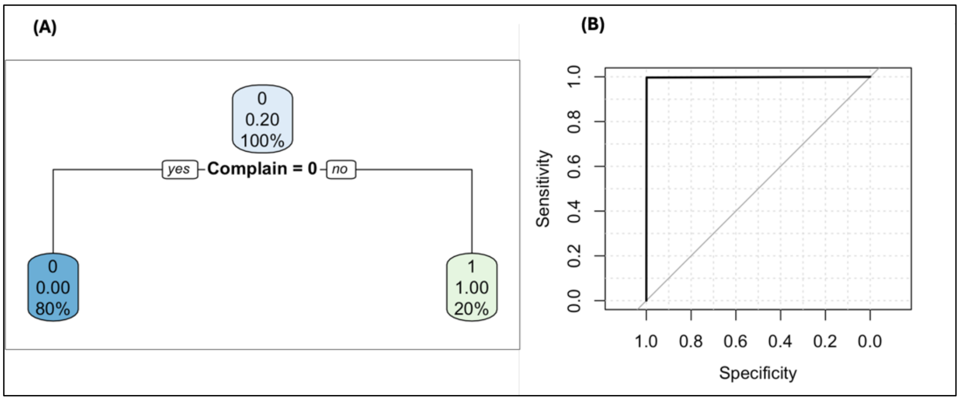


Figure 3 Decision Tree Results (A) Decision Tree Model (B) ROC curve of training results.

Classification of the test data set

Amongst all, the models chosen to be applied to the test set were (i) Single variable model from regression using "Complain" feature as the independent variable and (ii) Decision Tree classification. Both the models gave a test balanced accuracy of 99%, which is coherent with the balanced accuracy of 99% on training data. Firstly, this indicates that the choice of balanced accuracy measure was correct, as it played out well on the test data. Secondly, it is indicative of the generalizability of the model, and predictive power of both. In terms of computational efficiency, both the models are equally efficient. In terms of classification potential on unknown data, both the models perform equally well and surpass the requirement of predicting 80% customer churn. However, the insights obtained from the results of the two models differ. Decision trees give the purity index and thus the significance of selected independent variables (Branches and parent nodes) in classification. However, regression can quantitatively access the impact of the independent variables on the increase or decrease of dependent variables, thus creating more room for predictive analyses. In addition, regression allows more control over the selection of variables and variable subsets.

Conclusions

In conclusion, this report focused on improving the prediction of customer churn in a bank by refining feature selection and testing various classification models. It builds on the seven features identified in the previous work as essential for classification due to significant discrimination power, namely; Age, Balance, Gender, Geography, Number of Products, Is Active Member and Complaints. However, it was noted that the "Balance" feature was zero inflated and some customers who had exited the bank still maintaining large balances. Despite these issues, "Balance" was retained for analysis, while the "Is Active Member" variable was excluded due to inconsistencies. For evaluation of classification models, balanced accuracy measure was used due to the imbalance resultant class.

For classification, two models were explored using the k-Nearest Neighbors (kNN) algorithm. Model-1 utilized only quantitative variables, providing insights into the classification performance without the complexity of categorical features, acting as a base model. However, the results showed low classification accuracy (up to 46%), indicating the need for additional features. Model-2 incorporated mixed variables using Gower distances for dissimilarity calculation. This model yielded improved accuracy (up to 99%), demonstrating the importance of including categorical variables in classification. However, one of the strong drawbacks of this model is computational cost, significantly decreasing its efficiency. Second is the black box nature and complexity of understanding dimensions created from Gower's distance, and classification boundaries in multidimensional data created by kNN models. This makes the model difficult to interpret and manipulate, as well as complex to reproduce, whilst giving no insight into the weight and significance of variables on the resultant variable. Therefore, kNN will not be recommended as the first choice in classifiers.

Next, regression classifiers were employed, all utilizing binomial general linear models since our resultant variable is binomial in distribution. Regression Model-1 utilized all six features and Regression Model-2 employed subset selection to identify the most significant variables. The results revealed that a simpler model with just three variables—age, number of products, and complaint status—achieved comparable accuracy to the more complex model. Additionally, the LASSO method was used to further refine the model forming Model-3, ultimately identifying "Complain" as the sole significant variable for classification, giving a prediction balanced accuracy up to 99%.

Finally, a decision tree classifier was employed, offering simplicity and interpretability. The model effectively classified the data based on the "Complain" feature, demonstrating its ability to select features automatically and predict customer churn with high accuracy (up to 99%), without any complex parameter adjustments for a customer churn data set with strong discriminative features, and lacking many complexities.

The one variable regression and decision tree classification models were applied to a test dataset, yielding consistent balanced accuracy scores of 99% across both models. This indicates the generalizability and predictive power of the models. While both regression and decision tree models performed well, they offer different insights: decision trees provide clarity on feature importance, while regression allows for quantitative analysis of feature impact.

In conclusion, by addressing feature anomalies and selecting essential variables, the models were able to accurately predict customer churn. Further, the removal of less significant features improved model transparency, interpretability, and real-world applicability, while maintaining classification accuracy. A simpler model holds greater value in real life scenarios by decreasing data management and maintenance costs, as well as big data associated complexities. Overall, the one variable model from regression and decision tree both was the most efficient, appropriately accurate and realistically applicable. In this dataset, just following whether the customer has filed a complaint was enough to identify customer churn with up to 99% accuracy, crossing our required benchmark of predicting at least 80% clients who will churn. This approach emphasizes the importance of thoughtful feature selection and model refinement in building effective predictive models for customer churn, rather than resorting to very large and complex models before testing simpler ones. The simple and interpretable nature of one variables classifier make them highly stable, generalizable, parsimonious, and attractive in run time complexity. Further the nature of regression and decision tree classifiers make them simple and transparent enough to present and explain to stakeholders, while cross validation enables them to be stable enough to be reproducible.

In the future, customer churn prediction can be further improved upon by investigating trends in customer churn over time. Analyzing churn rates by month or quarter may reveal seasonal patterns or changes in customer behavior over time, which could inform targeted retention strategies. Segmenting customers based on demographic characteristics can help tailor the retention strategies to subgroups. Assessing model robustness to variations in input data or model parameters could provide prove model stability and generalizability.