Preliminary Results

Introduction

This project explores the use of predictive modeling to help the ABC Corporation identify customers at high risk of attrition. In the Preliminary Results , I present the development and evaluation of initial machine learning models, following the steps outlined in the Analytic Plan. The performance of these models will be compared, and areas for improvement will be identified and incorporated into the next steps for the Final Report.

1 Summary of Analytic Plan

The Analytic Plan outlined the initial steps for identifying customers at high risk of attrition for ABC Corporation. Guided by the business objective of retaining valuable customers and reducing acquisition costs, the plan included comprehensive data exploration, proposed feature engineering, and initial variable selection.

Throughout the initial exploratory analysis, both demographic and behavioral variables were examined to uncover attrition patterns. Key predictors that emerged as potentially informative included customer age, education level, transaction activity, credit usage, and inactivity periods. To refine the feature set, correlation analysis was used, and logarithmic transformations of skewed variables were proposed as part of the data preparation process.

Two supervised learning algorithms - Logistic Regression and k-Nearest Neighbors (kNN) - were selected for initial modeling due to their strengths in binary classification problems. Class imbalance, a critical issue in this dataset, undersampling techniques were proposed, particularly during kNN model training. Proposed evaluation strategies focused on using metrics such as accuracy AUC, and confusion matrices, with cross-validation applied to assess model generalizability. Ultimately, both models showed strong performance, with kNN achieving the highest recall and F1 score, while Logistic Regression provided clearer interpretability for business stakeholders.

The following sections detail the data preparation, model implementation, performance of the preliminary models, and highlight opportunities for improvement to address in the final report.

2 Data Preparation

2.1 Cleaning, Encoding, and Transformations

As outlined in the Analytic Plan, the data underwent several preprocessing steps including encoding, transformation, and verification for missing values. No missing (NA) values were found in the dataset before or after transformation.

2.1.1 Encoding Indicators

To allow for modeling flexibility, ordinal encodings were created for the categorical variables ‘Education_Level’ and ‘Card_Category’, with the following logic:

- Higher values represent more advanced education levels
- Increasing values for card categories corresponding to more premium card types

The encoding logic is detailed in Table 1 and Table 2 in the Appendix.

2.1.2 Variable Transformations

Logarithmic transformations were applied to skewed variables to address non-linearity. These transformations reduce skew and help linear models better capture relationships by stabilizing variance and compressing outliers. Transformations were performed on the following predictors:

- Credit_Limit

- Avg_Open_To_Buy
- Avg_Utilization_Ratio

A small constant was added to Avg_Utilization_Ratio to avoid errors resulting from taking the log of zero. The constant added was $min(x[x > 0])/2$, where x is the value of Avg_Utilization_Ratio. This conservative, data-driven value ensures minimal distortion while enabling valid transformation of this variable.

Visual inspections (shown in Figure 1, Figure 2, and Figure 3 in the Appendix) confirm that while transformations improved linearity and symmetry, some residual skew and non-linearity remained. For example, some curvature remains between 'Avg_Open_To_Buy' versus 'Avg_Utilization_Ratio' and the histogram of 'Avg_Utilization_Ratio' has a large number of observations between -7 and -6, with the remaining observations skewed to the left.

To support k-Nearest Neighbors (kNN) method, a standardized version of the dataset ('st_customers') was created with all numerical variables scaled (excluding the identifier column and response variable).

2.1.3 Robust Feature Selection

Thirteen predictors were initially selected based on exploratory analysis. These predictors were refined using the Lasso method via variable selection with the binomial family (appropriate for binary classification). Using 10-fold cross-validation:

- At lambda.min (value minimizing the cross-validation error), all variables were retained.
- At lambda.1se (the largest lambda within one standard error of lambda.min), Education_Level and Credit_Limit were dropped

The optimal values of lambda used as well as the predictors retained are shown in Table 3 in the Appendix.

The elimination of variables such as Education_Level and Credit_Limit, while surprising, suggests that they do not provide independent predictive value when other features (i.e., Avg_Utilization_Ratio and Total_Trans_Ct) are considered. This underscores the importance of transaction-based behaviors over demographic variables in predicting customer attrition. This insight could be influential in the creation of retention strategies and target marketing at ABC Corporation.

Lasso is particularly valuable here because it performs automatic feature selection by setting unimportant variable coefficients to zero. This not only simplifies the model but also supports the bias-variance trade-off by reducing the risk of overfitting, especially in high-dimensional data. Lasso's ability to enhance interpretability and eliminate noise variables make it a strong choice for predictor selection in this analysis.

3 Model Implementation

3.1 Data Balancing

To prepare the data for model implementation, approximately 80% of the original dataset was allocated to the training set (8,101 observations) while the remaining 20% formed the testing set

(2,026). Within the training set, 83.85% of the observations were labeled 'Existing Customer', revealing a significant class imbalance.

To mitigate the class imbalance, undersampling was applied to the majority class, reducing the number of 'Existing Customer' records to match the number of 'Attrited Customer' records. The resulting balanced training set, referred to as 'balanced_training_set', contains 1,308 observations for each class. Balancing the classes helps to ensure that the model learns efficiently from both groups.

The testing set was left unchanged to retain the original class distribution (84.25% 'Existing Customer'), to simulate real-world conditions.

3.2 Revised Feature Selection

To further address the impact of class imbalance on feature selection, the Lasso was reapplied using the balanced training set. This approach ensures that both classes are equally represented during variable selection, reducing potential bias towards the majority class. Consistent with earlier findings, this revised Lasso analysis shrank the coefficients for 'Education_Level' and 'Credit_Limit' to zero, reinforcing their limited predictive value. The full set of updated feature selection results is presented in Table 4 in the Appendix.

3.3 Logistic Regression Model

Three Logistic Regression models were built:

- Model 1: Full selected feature set (13 predictors; see Table 4)
- Model 2: Reduced set from Lasso (Education_Level, Credit_Limit removed)
- Model 3: Only statistically significant variable retained (Customer_Age, Total_Revolving_Bal removed)

3.3.1 Key Metrics & Visualizations

Model 1 yielded several statistically significant coefficients. However, the variables Customer_Age, Education_Level_dummy, Credit_Limit_log, and Total_Revolving_Bal were not statistically significant, as shown in the coefficient summary (Table 5 in the Appendix).

When applied to the withheld test data, the model produced the following confusion matrix:

		True Customer status		
		Existing	Attrited	Total
Predicted Customer status	Existing	1434	54	1488
	Attrited	273	265	538
	Total	1707	319	2026

From this, the accuracy was calculated as 83.86%, indicating that the model correctly classified approximately 84% of the cases. To further evaluate the performance, the following metrics were computed:

- Precision = 0.4926: Of the customers predicted to attrite, 49.26% actually did.
- Recall (Sensitivity) = 0.8307: The model correctly identified 83.07% of customers who actually attrited.
- F1 Score = 0.6184: A balance between precision and recall, indicating moderate model performance, particularly in capturing attrited customers.

These results suggest that while the model is quite effective at identifying attrited customers (high recall), its precision is lower, meaning it also misclassifies a number of existing customers as attrited.

The ROC curve, shown in Figure 4, illustrates the model's ability to discriminate between classes. The Area Under the ROC Curve (AUC) is 0.9241, which shows that the model has excellent discriminant power.

Model 2 also yielded several statistically significant coefficients. Once again, the variables Customer_Age and Total_Revolving_Bal were not statistically significant, as shown in the coefficient summary (Table 6 in the Appendix).

When applied to the withheld test data, the model produced the following confusion matrix:

		True Customer status		
		Existing	Attrited	Total
Predicted Customer status	Existing	1437	54	1491
	Attrited	270	265	535
	Total	1707	319	2026

From this, the accuracy was calculated as 84.01%, indicating that this model also correctly classified approximately 84% of the cases. To further evaluate the performance, the following metrics were computed:

- Precision = 0.4953: Of the customers predicted to attrite, 49.53% actually did.
- Recall (Sensitivity) = 0.8307: The model correctly identified 83.07% of customers who actually attrited.
- F1 Score = 0.6206: A balance between precision and recall, indicating moderate model performance, particularly in capturing attrited customers.

These results suggest that while this model is similarly effective to the Model 1 at identifying attrited customers (high recall), its precision is lower, meaning it also misclassifies a number of existing customers as attrited.

The ROC curve, shown in Figure 5, illustrates the model's ability to discriminate between classes. The Area Under the ROC Curve (AUC) is 0.9239, which shows that this model also has excellent discriminant power.

Model 3 was created using only the statistically significant variables. The coefficient summary is shown in Table 7 in the Appendix.

When applied to the withheld test data, the model produced the following confusion matrix:

		True Customer status		
		Existing	Attrited	Total
Predicted Customer status	Existing	1434	53	1487
	Attrited	273	266	539
	Total	1707	319	2026

From this, the accuracy was calculated as 83.91%, indicating that this model also correctly classified approximately 84% of the cases. To further evaluate the performance, the following metrics were computed:

- Precision = 0.4935: Of the customers predicted to attrite, 49.35% actually did.
- Recall (Sensitivity) = 0.8339: The model correctly identified 83.39% of customers who actually attrited.
- F1 Score = 0.6200: A balance between precision and recall, indicating moderate model performance, particularly in capturing attrited customers.

The ROC curve, shown in Figure 6, illustrates the model's ability to discriminate between classes. The Area Under the ROC Curve (AUC) is 0.9234, which shows that this model also has excellent discriminant power.

3.3.2 Performance Comparison Analysis

A summary of the key performance metrics resulting for the three Logistic Regression models is provided in Table 8 in the Appendix, along with a combined ROC curve plot (Figure 7). Each model was evaluated on the same test to ensure comparability.

Although the performance metrics across the three models are very similar, Model 3 - using the reduced set of nine predictors - offers nearly identical AUC and F1 score values to the more complex models. Given this minimal performance trade-off, the simple model is preferred for several reasons:

- Interpretability: A model with fewer predictors is easier to explain, particularly to stakeholders who may not have a technical background.
- Generalizability: By reducing the number of predictors, the model is less prone to overfitting the training data and more likely to perform well on test data.

- Efficiency: A leaner model requires fewer inputs, which can rescue data collection and processing overhead in a production environment.

Despite having fewer predictors, Model 3 maintained nearly identical accuracy and AUC. This supports its use in production, where fewer variables can reduce costs and improve clarity for decision-makers. Given the minimal trade-off in performance and the benefit, Model 3 will be evaluated against the selected k-Nearest Neighbors (kNN) model.

3.4 k-Nearest Neighbors Model

k-Nearest Neighbors (kNN) algorithm is a non-parametric model that can be sensitive to high-dimensional spaces, especially when the number of observations is not substantially greater than the number of predictors. To mitigate the effects of the curse of dimensionality, the Reduced Feature Set (shown in Table 8) was used, ensuring that the number of observations (n) remains substantially larger than the number of predictors (p).

To avoid classification ties and to explore model performance across different neighborhood sizes, kNN models were trained using odd values of k ranging from 1 to 19.

3.4.1 Key Metrics & Visualizations

Table 9 summarizes the key metrics (Accuracy, F1 Score, and AUC) for each value of k. Accuracy and F1 score are maximized when $k = 11$, while AUC continues to increase slightly, reaching its maximum at $k = 19$. However, the AUC gain from $k = 11$ to $k = 19$ is minimal (0.24%), suggesting diminishing returns.

Figure 8 visualizes how each metric changes as k increases. The plot confirms that $k = 11$ is the optimal value for achieving balanced performance. While AUC continues to improve slightly beyond $k = 11$, the practical impact of these increases was minimal. The importance of identifying attrited customers without increasing the number of false positives makes the model at $k = 11$ the most balanced and interpretable version.

3.4.2 Performance Analysis for k = 11

Based on the analysis in the previous section, the kNN model for $k = 11$ neighbors was selected for final evaluation. The confusion matrix based on the withheld test data is shown below:

		True Customer status		
		Existing	Attrited	Total
Predicted Customer status	Existing	1514	47	1561
	Attrited	193	272	465
	Total	1707	319	2026

From this, the following performance were metrics were computed and interpreted:

- Accuracy = 0.8815: The model correctly classifies 88.15% of the test cases.
- Precision = 0.5849: Of the customers predicted to attrite, 58.49% actually did.
- Recall (Sensitivity) = 0.8527: The model correctly identified 85.27% of customers who actually attrited.
- F1 Score = 0.9265: A balance between precision and recall, indicating strong model performance.

These results indicate that the model is highly effective at capturing customers likely to attrite (high recall), but also includes some false positives (lower precision). The high F1 Score suggests that the model maintains a strong overall balance between Precision and Recall, making it suitable for identifying at-risk customers with reasonable confidence.

4 Model Comparison

4.1 Null Model

The null model based on the test data is included in the comparison to provide a baseline for evaluating the performance of the selected Logistic Regression model and kNN ($k = 11$) models. This model always predicts the majority class - in this case, "Existing Customer", which makes up 84.25% of the test data (1,707 out of 2,026). Thus, the null model achieves an accuracy rate of 84.25%. Because it fails to identify any attrited customers, the resulting Recall and F1 Score of 0.

4.2 Model Performance Comparison

A comparison between performance metrics of the null model, Logistic Regression model, and kNN ($k = 11$) can be seen in Table 10. Compared to the null model (baseline), both the Logistic Regression and kNN models offer substantial improvements in accuracy as well as their predictive accuracy for attrited customers. The kNN model in particular has a strong recall and F1 score, demonstrating meaningful predictive value beyond both the null model and the Logistic Regression model. The kNN model's superior recall likely stems from its ability to adapt to non-linear patterns, which may not follow the assumptions of the Logistic Regression model. However, the Logistic Regression model could be favorably based on model interpretation and simplicity, especially if an understanding of each predictor's influence is needed. While the kNN model does not differentiate between predictors' importance, it may be favored for strong predictive accuracy.

4.3 Why Model Performance Differs

Differences in the models result from the unique strengths and assumptions of each algorithm. Logistic Regression is a linear, parametric model that assumes a log-linear relationship between the predictors and the probability of customer attrition. The benefits of the Logistic Regression model include simplicity and interpretability.

On the other hand, kNN is a non-parametric, instance-based algorithm that makes no assumptions about the underlying data distribution. When the decision line is highly non-linear

and n>>p, kNN will dominate the logistic regression model. However, kNN is more sensitive to feature scaling and noise, and does not allow for easy interpretability since it does not differentiate between the importance of predictors.

These distinctions between the Logistic Regression model and the kNN model explain the higher values for recall and F1 score achieved by the kNN model - the flexibility of the kNN model helps to detect patterns associated with customer attrition. While the Logistic Regression model is slightly less accurate, it offers interpretability and transparency, which can help justify and explain business decisions more clearly. Additionally, the Logistic Regression model's coefficients easily trace back to business factors (e.g., higher transaction frequency reducing attrition risk), making it more suitable for scenario planning or policy development.

5 Next Steps

5.1 Expanded Exploratory Data Analysis (EDA)

To use more involved variable selection and enhance model performance, more extensive EDA will need to occur in the next phase of the project. To detect interactions between variables, facet plots will be reexamined to visualize how pairs of variables together relate to attrition. Boxplots grouped by qualitative variables and colored by attrition will be examined as well. Based on the results, the effectiveness of interactions will be analyzed using Lasso to determine which have the most predictive power on customer attrition.

To examine multicollinearity between variables, a heat map for correlation coefficients will be created. If multicollinearity is detected, this will help determine the appropriateness of the Logistic Regression model fit in the Preliminary Results. The presence of multicollinearity can decrease the interpretability of coefficients and may make predictors that matter appear statistically insignificant.

5.2 Further Feature Engineering

New features that will be calculated and introduced in the final report to fine tune predictive models include:

- Average Transaction Size: Can be used to determine unusual spending patterns - i.e., customers who make infrequent but large purchases versus those who are making frequent small purchases
- Engagement Score: Combination of multiple activity predictors (total transaction counts, amount, and change between Q1 and Q4) in relation to credit limit
- Transaction Count to Relationship Count: Ratio used to indicate the depth of customer relationship

5.3 Ensemble Methods

Decision trees can be advantageous over approaches such as Logistic Regression and kNN for classification models. They can handle qualitative predictors without the need for dummy

variables and are less technical to interpret. To capitalize on the opportunity to use an ensemble method involving decision trees, the Random Forest Method and Gradient Boosting Machines (GBM) will be utilized in the Final Project. Random forest and GBM models can handle more complex situations such as non-linear relationships and variable interactions better than Logistic Regression models. The downside is their higher computational cost, but it can be a reasonable price to ensure improved precision in the model performance.

In summary, the Final Report will aim to deliver:

- An ensemble model with superior precision, recall, and F1 score
- Deeper behavior insights via new engineered features
- Model outputs tailored for actionable business interventions.

6 Conclusion

The Preliminary Results successfully implemented and evaluated the Logistic Regression and kNN models for predicting customer attrition at ABC Corporation. Despite class imbalance and feature complexity, both models were highly accurate, achieving high performance. The kNN model could have the upper hand due to its superior predictive power, yet the greater interpretability of the Logistic Regression model may make it more desirable given the situation. In the Final Report, enhancing model complexity through ensemble methods and refined features is recommended.

7 Appendix

Table 1

VARIABLE: 'Education_Level'	
Numeric Indicators	***Category***
0	'Unknown'
1	'Uneducated'
2	'High School'
3	'College'
4	'Graduate'
5	'Post-Graduate'
6	'Doctorate

Table 2

VARIABLE: 'Card_Category'	
Numeric Indicators	***Category***
0	'Blue'
1	'Silver'
2	'Gold'
3	'Platinum'

Table 3

Lasso Method Results		
λ ***Value***	***Variable Set #***	***Resulting Predictors***
$\lambda = 0.00022$ *based on lowest cross-validation error,* *predictors from initial variable selection*	1	● 'Customer_Age' ● 'Education_Level' *(dummy)* ● 'Total_Relationship_Count' ● 'Card_Category' *(dummy)* ● 'Months_Inactive_12_mon' ● 'Credit_Limit' *(transformed)* ● 'Total_Revolving_Bal' ● 'Avg_Open_To_Buy' *(transformed)* ● 'Total_Amt_Chng_Q4_Q1' ● 'Total_Trans_Amt', ● 'Total_Trans_Ct' ● 'Total_Ct_Chng_Q4_Q1' ● 'Avg_Utilization_Ratio' *(transformed)*
$\lambda = 0.00526$ *based on lowest cross-validation error, all predictors*	2	● 'Customer_Age' ● 'Total_Relationship_Count' ● 'Card_Category' *(dummy)* ● 'Months_Inactive_12_mon' ● 'Total_Revolving_Bal' ● 'Avg_Open_To_Buy' *(transformed)* ● 'Total_Amt_Chng_Q4_Q1' ● 'Total_Trans_Amt', ● 'Total_Trans_Ct' ● 'Total_Ct_Chng_Q4_Q1' ● 'Avg_Utilization_Ratio' *(transformed)*

Table 4

Re-Applied Lasso Method Results		
λ ***Value***	***Variable Set #***	***Resulting Predictors***
$\lambda = 0.00080$ *based on lowest cross-validation error,* *predictors from initial variable selection*	3	● 'Customer_Age' ● 'Education_Level' *(dummy)* ● 'Total_Relationship_Count' ● 'Card_Category' *(dummy)* ● 'Months_Inactive_12_mon' ● 'Credit_Limit' *(transformed)* ● 'Total_Revolving_Bal' ● 'Avg_Open_To_Buy' *(transformed)* ● 'Total_Amt_Chng_Q4_Q1' ● 'Total_Trans_Amt', ● 'Total_Trans_Ct' ● 'Total_Ct_Chng_Q4_Q1' ● 'Avg_Utilization_Ratio' *(transformed)*
$\lambda = 0.00986$ *based on cross-validation error within one standard error of the minimum,* *predictors from initial variable selection*	4	● 'Customer_Age' ● 'Total_Relationship_Count' ● 'Card_Category' *(dummy)* ● 'Months_Inactive_12_mon' ● 'Total_Revolving_Bal' ● 'Avg_Open_To_Buy' *(transformed)* ● 'Total_Amt_Chng_Q4_Q1' ● 'Total_Trans_Amt', ● 'Total_Trans_Ct' ● 'Total_Ct_Chng_Q4_Q1' ● 'Avg_Utilization_Ratio' *(transformed)*

Table 5

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	7.953e+00	8.566e-01	9.285	< 2e-16	***
Customer_Age	-9.988e-03	7.5033-03	-1.331	0.18313	
Education_Level_dummy	8.585e-03	3.375e-02	0.254	0.79921	
Total_Relationship_Count	-3.389e-01	3.960e-02	-8.559	< 2e-16	***
Card_Category_dummy	5.730e-01	1.907e-01	3.005	0.00265	**
Months_Inactive_12_mon	4.865e-01	5.991e-02	8.121	4.63e-16	***
Credit_Limit_log	3.236e-01	2.079e-01	1.556	0.11966	
Total_Revolving_Bal	-8.134e-06	1.366e-04	-0.0606	0.95252	
Avg_Open_To_Buy_log	-6.265e-01	1.549e-01	-4.044	5.26e-05	***
Total_Amt_Chng_Q4_Q1	-6.663e-01	2.803e-01	-2.377	0.01746	*
Total_Trans_Amt	5.395e-04	3.699e-04	14.587	< 2e-16	***
Total_Trans_Ct	-1.258e-01	5.930e-03	-21.217	< 2e-16	***
Total_Ct_Chng_Q4_Q1	-2.797e+00	2.738e-01	-10.217	< 2e-16	***
Avg_Utilization_Ratio_log	-4.842e-01	5.276e-02	-9.177	< 2e-16	***

Signif. Codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 3626.5 on 2615 degrees of freedom
Residual deviance: 1869.8 on 2602 degrees of freedom
AIC: 1897.8

Number of Fisher Scoring iterations: 6

Table 6

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	8.812e+00	6.676e-01	13.200	< 2e-16	***
Customer_Age	-9.800e-03	7.485e-03	-1.309	0.190441	
Total_Relationship_Count	-3.400e-01	3.956e-02	-8.593	< 2e-16	***
Card_Category_dummy	6.237e-01	1.880e-01	3.318	0.000907	***
Months_Inactive_12_mon	4.859e-01	5.988e-02	8.115	4.85e-16	***
Total_Revolving_Bal	7.743e-05	1.248e-04	0.620	0.535051	
Avg_Open_To_Buy_log	-4.044e-01	5.712e-02	-7.080	1.44e-12	***
Total_Amt_Chng_Q4_Q1	-6.611e-01	2.798e-01	-2.363	0.018116	*
Total_Trans_Amt	5.406e-04	3.689e-05	14.655	< 2e-16	***
Total_Trans_Ct	-1.261e-01	5.930e-03	-21.260	< 2e-16	***
Total_Ct_Chng_Q4_Q1	-2.795e+00	2.734e-01	-10.221	< 2e-16	***
Avg_Utilization_Ratio_log	-4.810e-01	5.276e-02	-9.177	< 2e-16	***

Signif. Codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 3626.5 on 2615 degrees of freedom
Residual deviance: 1872.3 on 2604 degrees of freedom
AIC: 1896.3

Number of Fisher Scoring iterations: 6

Table 7

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	8.470e+00	5.701e-01	14.585	< 2e-16	***
Total_Relationship_Count	-3.985e-01	3.952e-02	-8.590	< 2e-16	***
Card_Category_dummy	6.322e-01	1.883e-01	3.358	0.000785	***
Months_Inactive_12_mon	4.763e-01	5.954e-02	8.000	1.25e-15	***
Avg_Open_To_Buy_log	-3.993e-01	5.421e-02	-7.366	1.76e-13	***
Total_Amt_Chng_Q4_Q1	-6.524e-01	2.792e-01	-2.336	0.019471	*
Total_Trans_Amt	5.424e-04	3.686e-05	14.741	< 2e-16	***
Total_Trans_Ct	-1.258e-01	5.911e-03	-21.290	< 2e-16	***
Total_Ct_Chng_Q4_Q1	-2.802e+00	2.723e-01	-10.292	< 2e-16	***
Avg_Utilization_Ratio_log	-4.540e-01	2.867e-02	-15.835	< 2e-16	***

Signif. Codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 3626.5 on 2615 degrees of freedom
Residual deviance: 1874.5 on 2606 degrees of freedom
AIC: 1894.5

Number of Fisher Scoring iterations: 6

Table 8

Variable Set	Accuracy	Precision	Recall	F1 Score	AUC
Variable Set 1 in Table 3	0.8386	0.4926	0.8307	0.6184	0.9241
Variable Set 2 in Table 3	0.8401	0.4953	0.8307	0.6206	0.9239
Reduced Feature Set*	0.8391	0.4953	0.8339	0.6200	0.9234

* The reduced model includes the following predictors: Total_Relationship_Count, Card_Category *(dummy)*, Months_Inactive_12_mon, Avg_Open_To_Buy *(transformed)*, Total_Amt_Chng_Q4_Q1, Total_Trans_Amt, Total_Trans_Ct, Total_Ct_Chng_Q4_Q1, Avg_Utilization_Ratio *(transformed)*.

Table 9

k	Accuracy	F1	AUC
1	0.8633	0.9148	0.8449
3	0.8810	0.9263	0.9184
5	0.8796	0.9252	0.9314
7	0.8776	0.9240	0.9360
9	0.8791	0.9250	0.9391
11	0.8815	0.9265	0.9410
13	0.8771	0.9236	0.9408
15	0.8727	0.9207	0.9421
17	0.8717	0.9200	0.9431
19	0.8707	0.9194	0.9434

Table 10

Metric	Null Model	Logistic Regression	kNN (k=11)
Accuracy	0.8425	0.8401	0.8815
Precision	-	0.4953	0.5849
Recall	0.0000	0.8307	0.8527
F1 Score	0.0000	0.6206	0.9265
AUC	0.5000	0.9239	0.9410

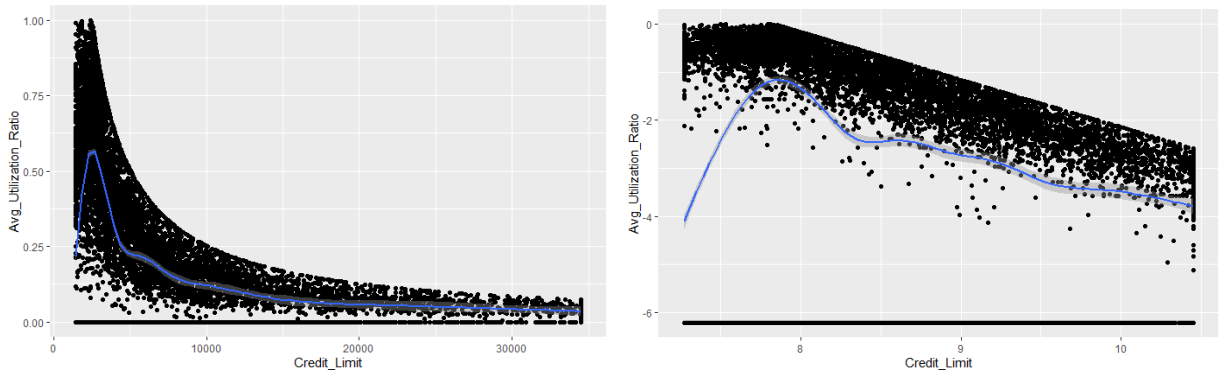

Figure 1 *shows the relationship between 'Credit_Limit' and 'Avg_Utilization_Ratio' before performing the logarithmic transformation to both variables* [left] *and after performing the logarithmic transformation to both variables* [right].

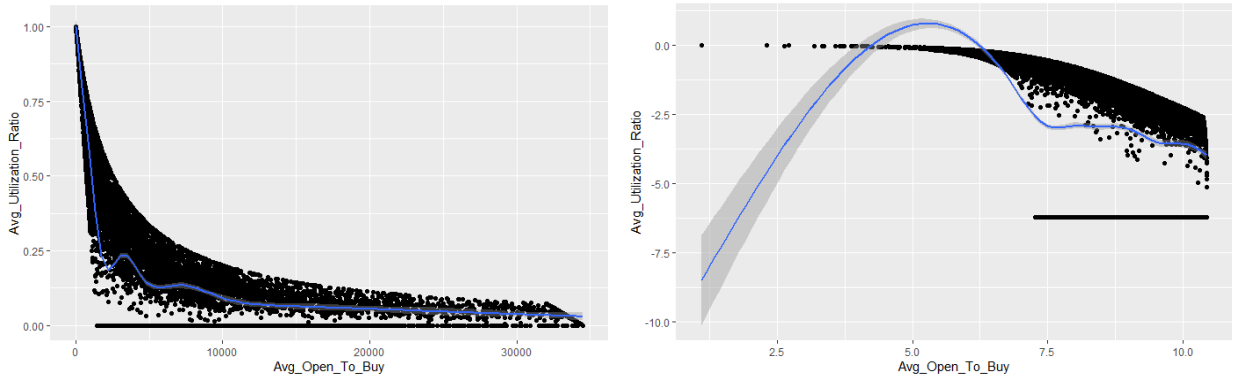

Figure 2 *shows the relationship between 'Avg_Open_To_Buy' and 'Avg_Utilization_Ratio' before performing the logarithmic transformation to both variables* [left] *and after performing the logarithmic transformation to both variables* [right].

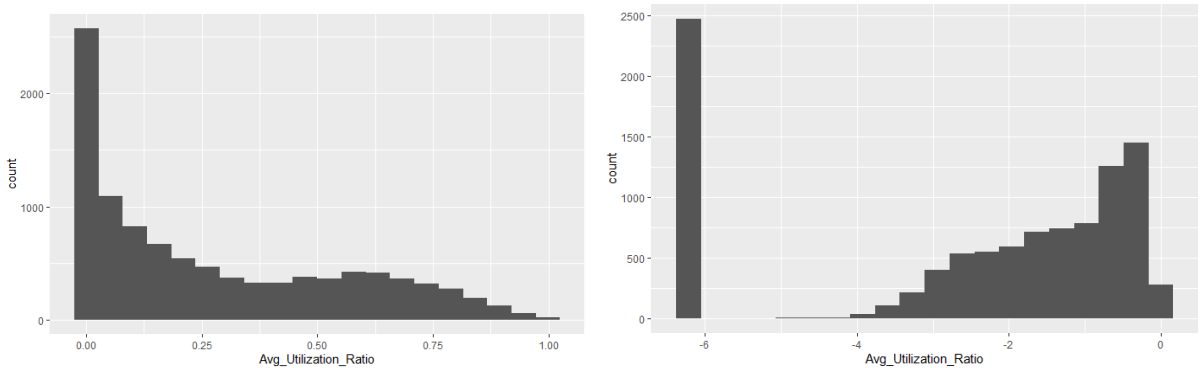

Figure 3 *shows the distribution of'Avg_Utilization_Ratio' before performing the logarithmic transformation* [left] *and after performing the logarithmic transformation* [right].

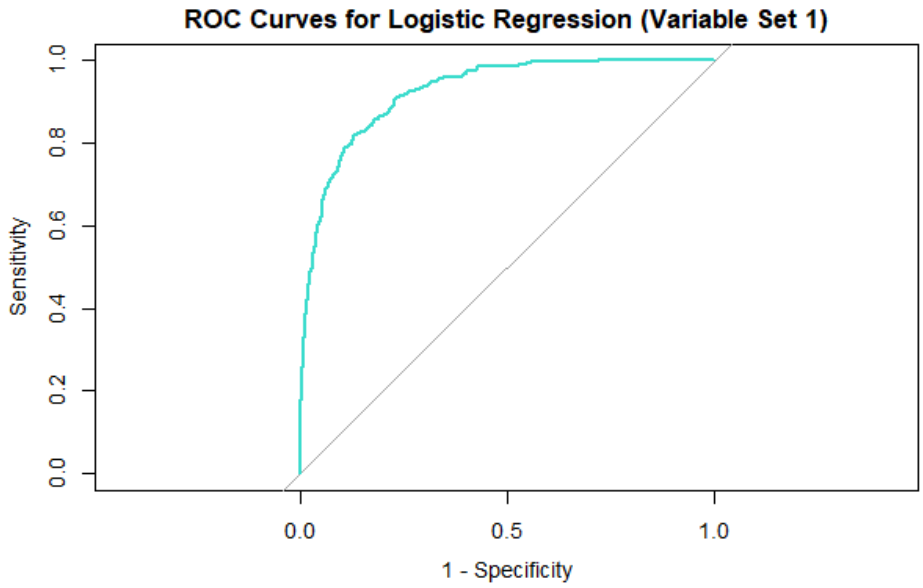

Figure 4 *shows the ROC Curve for the Logistic Regression model fit from Variable Set 1.*

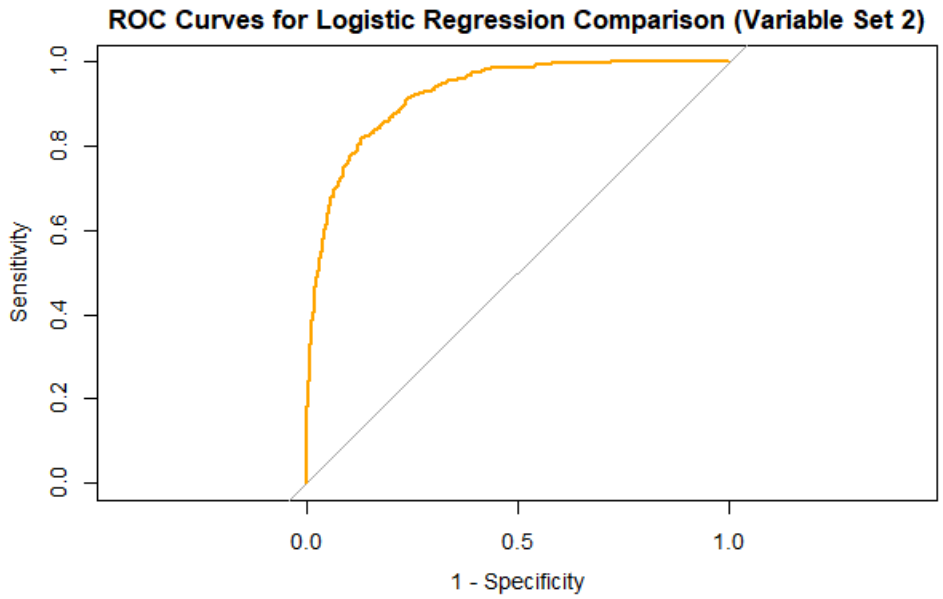

Figure 5 *shows the ROC Curve for the Logistic Regression model fit from Variable Set 2.*

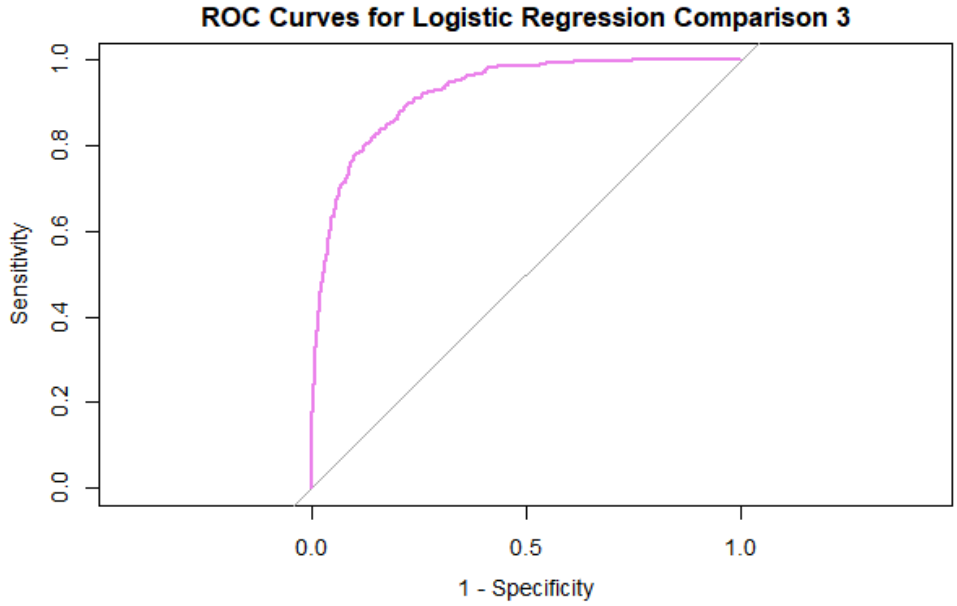

Figure 6 *shows the ROC Curve for the Logistic Regression model fit from the statistically significant variables.*

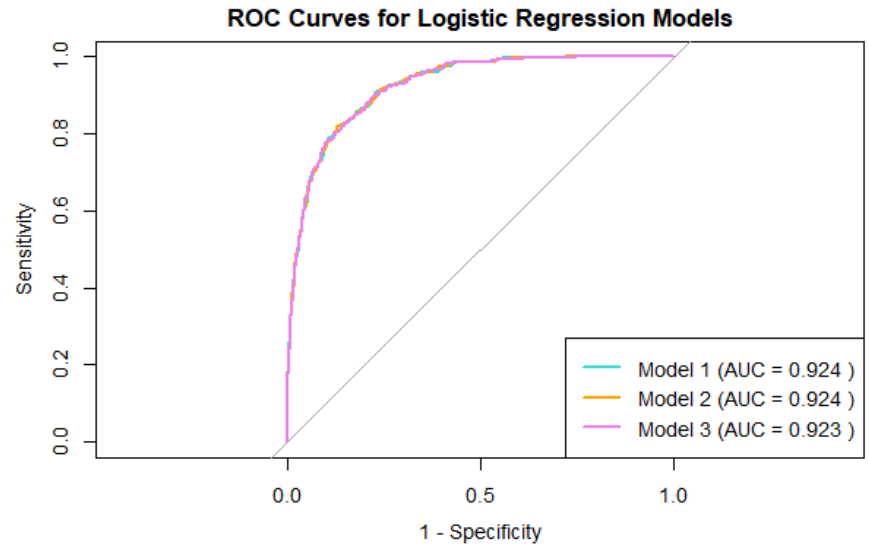

Figure 7 *shows the ROC Curve for the all three Logistic Regression models.*

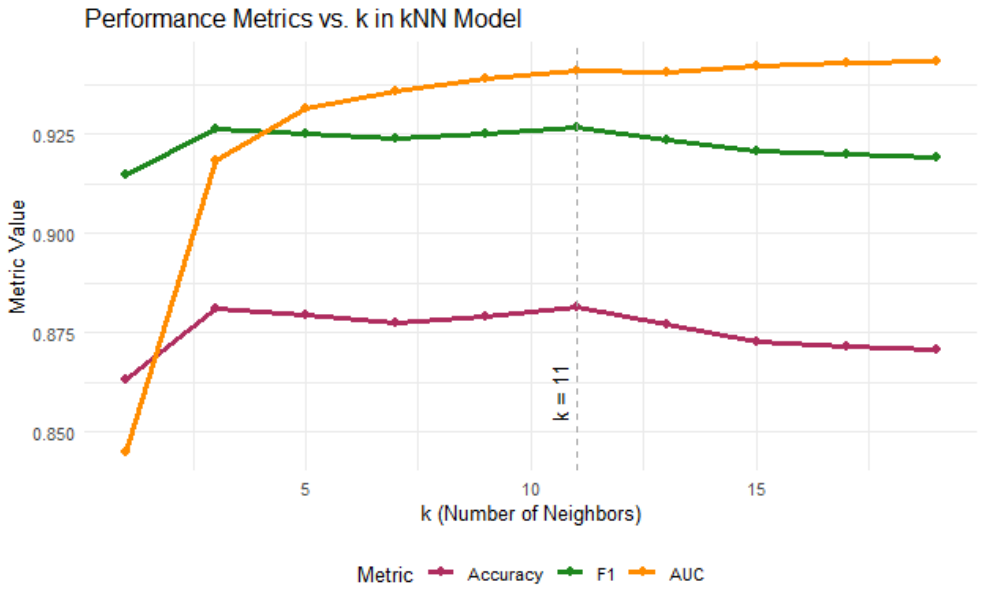

Figure 8 *tracks how each metric behaves as the number of neighbors (k) increases; the dashed line at k = 11 reinforces the selection of the kNN model when k=11.*

8 References

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

The Pecan Team. (2024, August 5). *Top ML Models for Prediction Customer Churn: A Comparative Analysis*. https://www.pecan.ai/blog/best-ml-models-for-predicting-customer-churn/