Table of Contents

Section I: Introduction

1.1 Background

This is a 'vintage' of credit card holders that has been randomly sampled. There are several fields in this data. The dependent variable is named 'bad' with outputs 0 and 1 where 0=the person is a good customer and 1=the person is a bad customer. We have several independent variables like opening balance, ending balance, days of delinquency, over limit amount, city, State and so on.

When a customer becomes bad, the variable named bad will have a value of 1. Its value otherwise will be 0. What constitutes a bad customer is up for debate. However, in this data we have monthly 'performance' data from the customers to include their purchases, payments, account status, days delinquent, etc. The timing of events is EXTREMELY important in 'predictive' modeling of transactional style data. Generally speaking, the inputs/predictors are not measured in the same timeframe as the outcomes/dependent variable. In this example, our function is to predict the outcome of the customer based on their status at the end of their third month as a customer (MOB=3). Those values at month 3 become the inputs for prediction. For purposes of this problem, we will define bad based on their last observed status as long as that status is after MOB=3. If the customer in the last observation of their account has DaysDeliq>=60 or the account has an 'ExternalStatus' that is anything other than blank or 'C' then the customer has gone bad and bad = 1.

Additionally, there are 2 variables that reflect profit indirectly. We do not have 'profit' but payments and purchases are 'directionally' related to profit. Sum of the TotalNetPayments and TotalNetPurchaseAndCash for each customer for the duration of the observations (sum them across time) are two such variables. Similarly, when an account is 'bad' it will default on its debt. The balance on that

last statement is considered a loss. We have created a field called ‘chargeoffdollars’ from this information and summarized.

1.2 Objectives

The objective of the study is to evaluate the performance of two predictive models, logistic regression and decision tree, in predicting whether a customer will go bad or not. The study aims to determine which model is more effective in predicting customer risk and to identify the most important features that contribute to a customer going bad.

1.3 Exploratory Data Analysis

All statistical analysis was conducted in R and RStudio and plots were created using ggplots2. Our goal is to select variables that best predict our chosen dataset.

Table 1. Bad Rate Summary

Bad clients	Frequency
NO	3122
YES	4453
TOTAL	7575

From this table, out of the 7575 customers, 4453 are bad while 3122 clients are good clients. Hence, about 58% of the customers are bad.

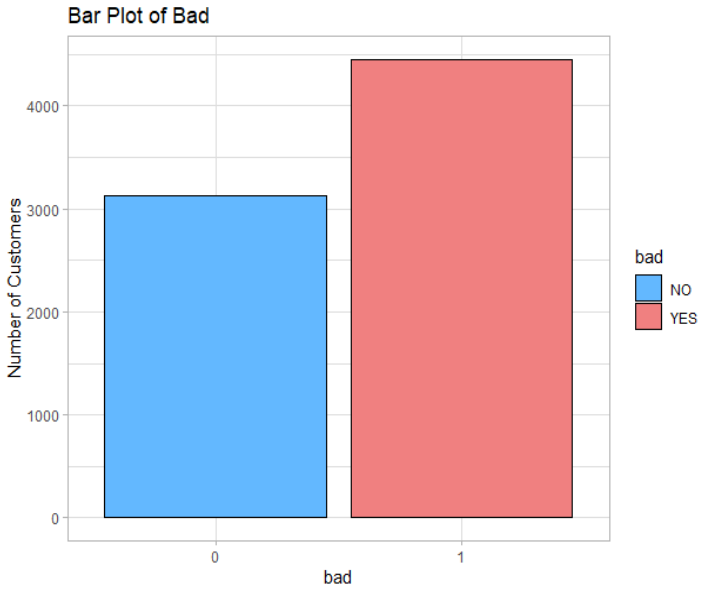

Figure 1. Bar plot of bad and good customers

This figure shows the number of bad and good customers. We can see that number of bad customers are higher than that of good customers.

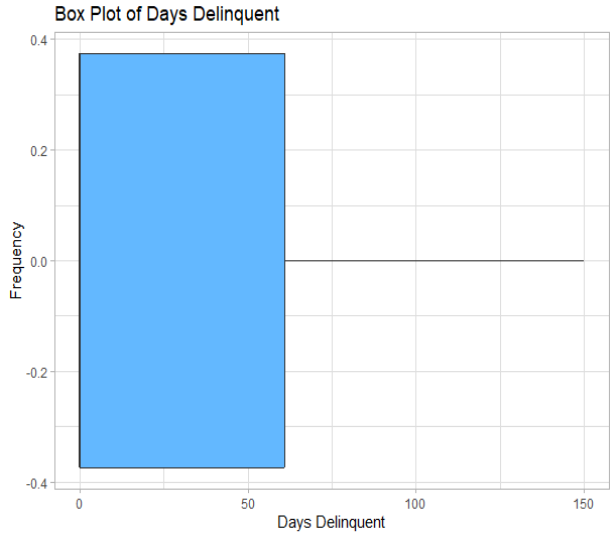

Figure 2. Box plot of days delinquent

- The minimum value of Days Delinquent is 0, meaning that some customers have never been delinquent.
- The first quartile (25th percentile) of Days Delinquent is also 0, indicating that 25% of customers have never been delinquent or have been delinquent for less than a day.
- The median value of Days Delinquent is 0, suggesting that most customers have never been delinquent or have been delinquent for a short period of time.
- The mean value of Days Delinquent is 33.63, which is greater than the median. This indicates that there are some customers who have been delinquent for a longer period, which is pulling the mean upwards.
- The third quartile (75th percentile) of Days Delinquent is 61, indicating that 75% of customers have never been delinquent or have been delinquent for less than 61 days.
- The maximum value of Days Delinquent is 150, suggesting that some customers have been delinquent for a significant period.

Table 2: Summary statistics for the lifetime purchase and lifetime payments

Bad clients / **Summary**	**Lifetime purchase**	**Lifetime payments**
Min	-57.5	-190.0
Median	541.0	528.8
Mean	744.4	734.0
Max	17551.9	18134.7

Table 2 shows the distribution of the lifetime purchase which has a minimum value of 57.5, mean value of 744.4 and the maximum value of 17551.9 while the lifetime payment has a minimum value of-190.0, mean value of 734.0 and the maximum value of 18134.7.

Table 3: Summary Statistics for the pertinent inputs

INPUTS Summary	MOB	DAYS Delinquent	Actual Min Pay	Over limit Amount	Credit Limit
Min	3	0	0	0	0
Median	3	0	48	0	300
Mean	3	33.63	70.73	29.31	415.5
Max	3	150	341	811.97	700

Table 3 shows the summary of some of the pertinent inputs in this study. The MOB has a minimum value of 3, median value of 3, mean value of 3 and the maximum value of 3. Days delinquent has a minimum value of 0, mean value of 33.63, and the maximum value of 150. However, the actual minimum pay has a minimum value of 0, mean value of 70.73 and the maximum value of 341. Also, over limit amount has a minimum value of 0, mean value of 29.31 and maximum value of 811.97 while the credit limit has a minimum value of 0, median value of 300, mean value of 415.5 and maximum value 700.

Table 4: Summary Statistics for the pertinent inputs continues.

INPUTS Summary	Display Min Pay	Opening Balance	Ending Balance	Quarterly Credit Score	Total Fees Billed	B score
Min	0	-441.2	-397.8	0	-277.17	234
Median	36.16	337.6	317.1	583	18.93	614
Mean	81.30	372.6	339.3	557.4	21.84	531.8
Max	921.97	1099.7	1235.2	791	190.52	695

Table 4 also shows the summary of some of the pertinent inputs in this study. The Display minimum pay has a minimum value of 0, median value of 36.16, mean value of 81.30 and the maximum value of 921.97. Opening balance has a minimum value of -441.2, mean value of 372.6 and maximum value of 1099.7. However, the ending balance has a minimum value of -397.8, mean value of 339.3 and the maximum value of 1235.2. Also, the quarterly credit score has a minimum value of 0, mean value of 557.4 and maximum value of 791 while the total fees have a minimum value of -277.17, median value of 18.93,

mean value of 21.84 and maximum value 190.52. And the B score has a minimum value of 234, median value of 614, mean value of 531.8 and the maximum value of 695.

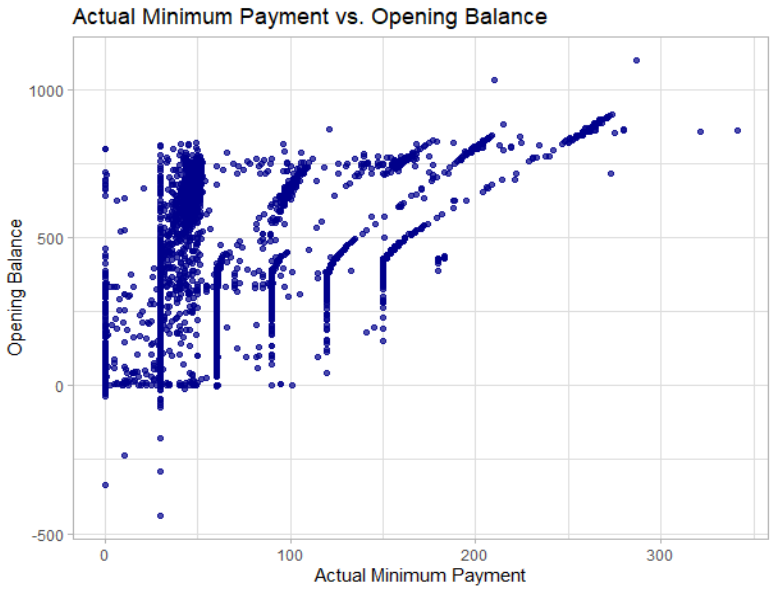

Figure 3. Scatterplot of Actual Minimum Payment VS Opening Balance

The plot suggests that as the opening balance increase, values of Actual Minimum Payment also tend to increase, though not necessarily in a perfectly linear fashion.

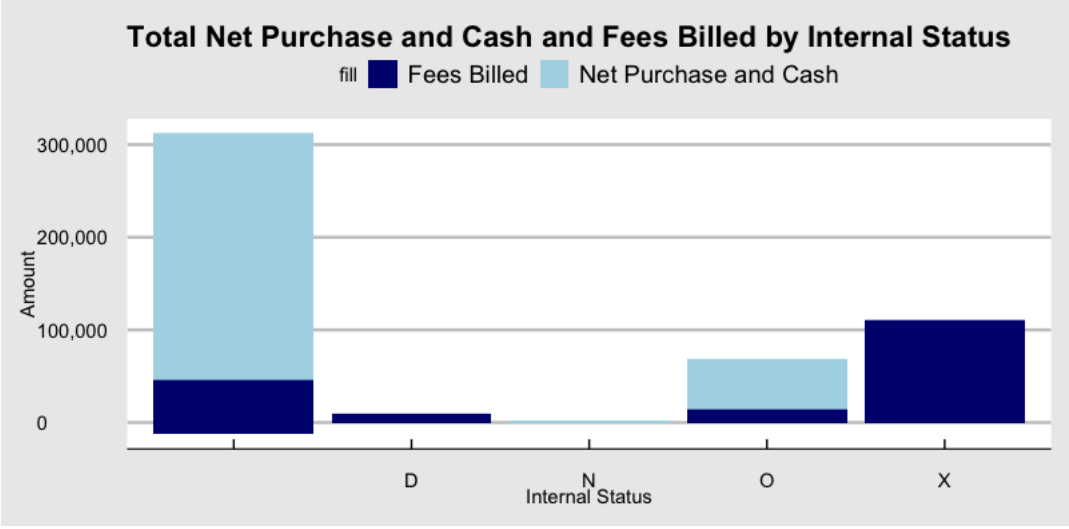

Figure 4. Total net purchase and cash and fees based on internal status.

It appears that the group with a blank Internal Status had the highest levels of Net Purchase and Cash compared to all other Internal Statuses. On the other hand, the group with internal status X had the highest amount fees billed. So, Internal Status X has a higher probability of being bad as they have a lot of bills piled up.

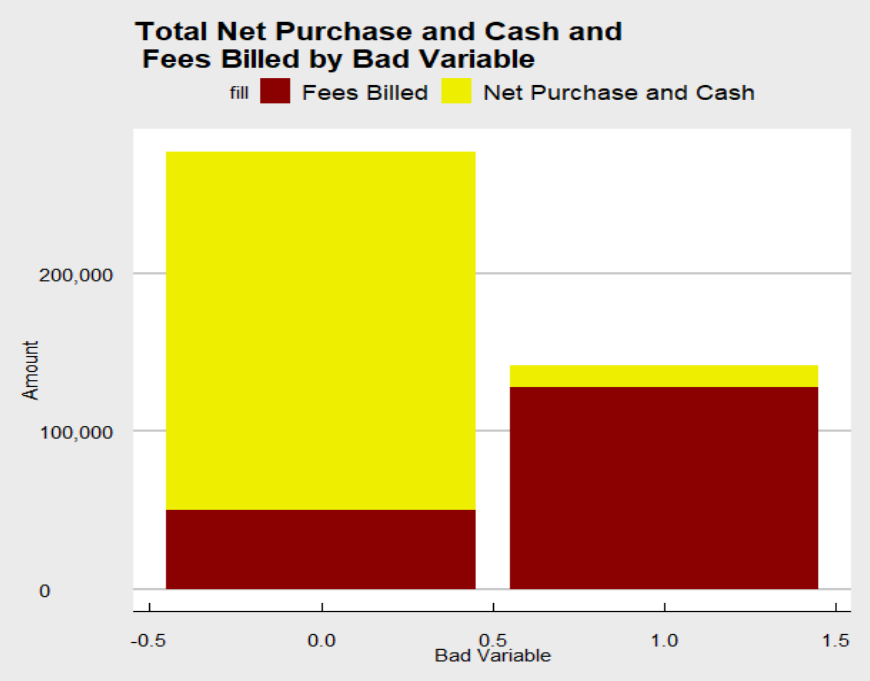

Figure 5. Total net purchase and cash and fees based on bad status

Here, it appears that the group with a good (not bad) Status had a higher level of Net Purchase and Cash compared to the bad group. On the other hand, the bad group had the highest amount fees billed. This verifies our previous assumption that people with a large amount of fees billed are more likely to become bad.

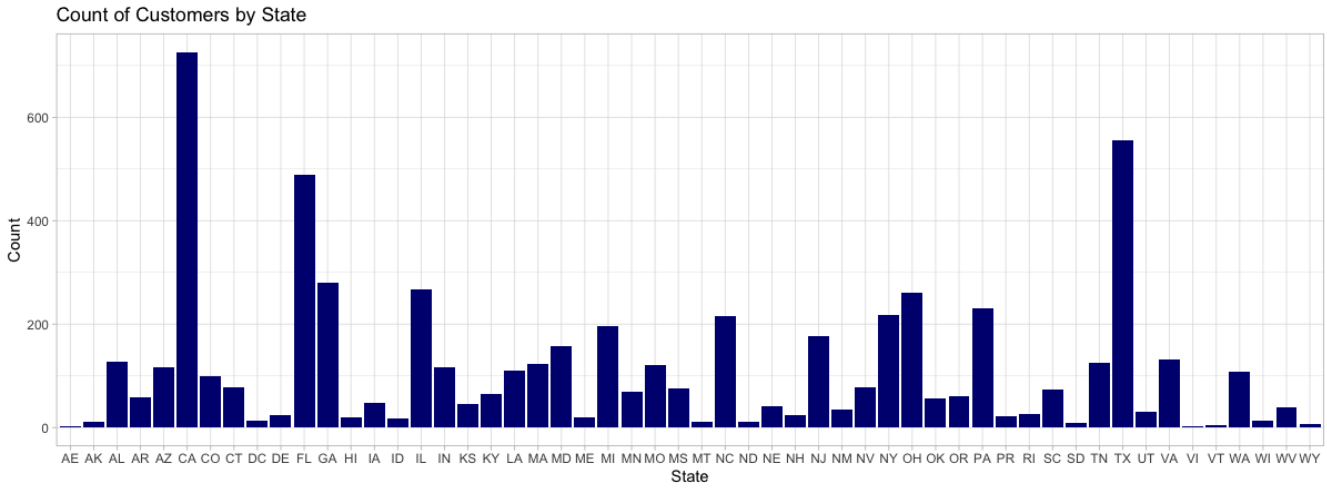

Figure 6. Count of customers by state

The graph provides evidence that California, Texas, and Florida had the largest customer bases compared to other states included in the dataset. This observation may indicate that these states have a larger population or a higher demand for the type of product or service being offered.

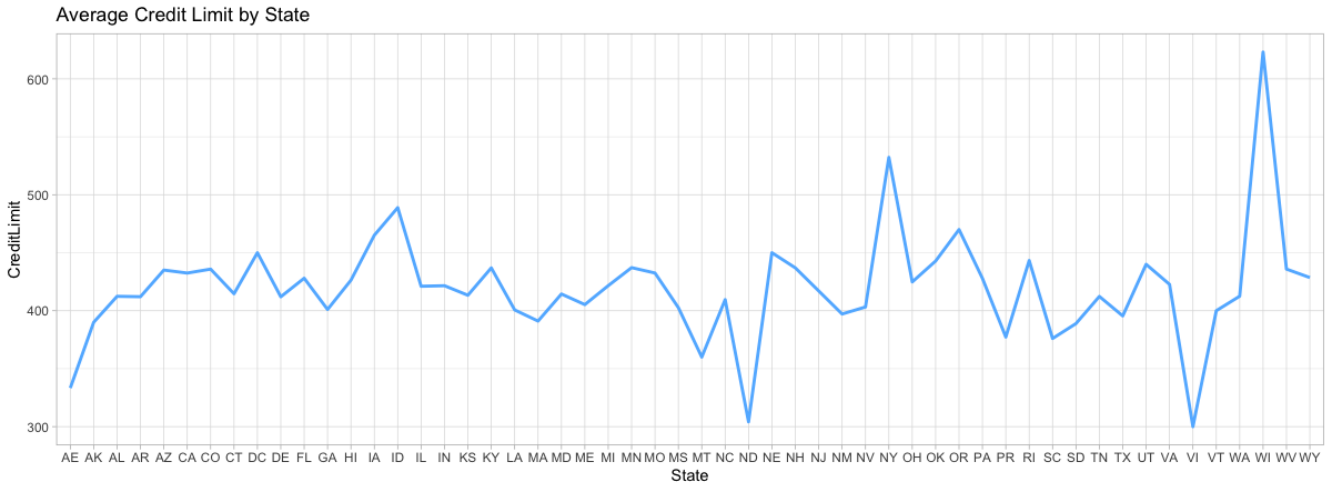

Figure 7. Average credit limit by state

The graph presented indicates that Wisconsin and New York had the highest average credit limit compared to other states, while the Virgin Islands and North Dakota had the lowest average credit limit. This observation could suggest that customers in Wisconsin and New York have a higher

creditworthiness or financial stability, while customers in the Virgin Islands and North Dakota may have lower creditworthiness or financial stability.

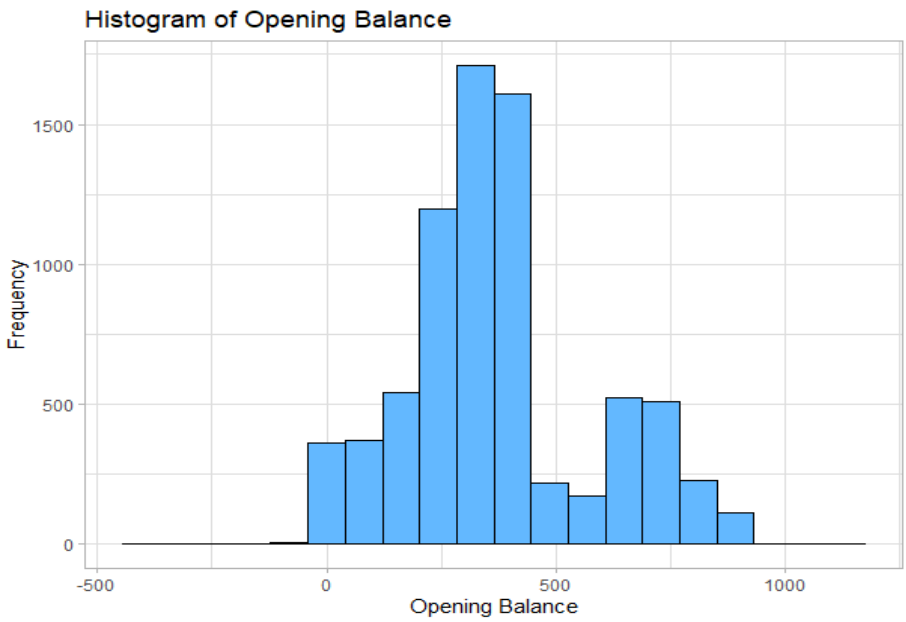

Figure 8. Histogram of opening balance

From the histogram we can see that many people have 300 dollars opening balance which is expected given the fact that majority of the customers have the credit limit of 300 dollars. We can see that none of the customers had spent more than 1000 dollars during their first month.

Section II: Model Fitting

2.1 Regression Trees

The tree model is a predictive model that uses a decision tree to classify data. In this report, we will discuss the tree model created using the rpart function in R, which aims to predict the probability of a client being bad based on their Bscore, quarterly credit score, overlimit amount, total fees billed, opening balance and other related variables.

The model uses the rpart algorithm to build a decision tree that identifies the most important predictors for predicting graduation. The predictors used in this model are ExtStatus, DaysDeliq, IntStatus, ActualMinPay, OverlimitAmount, DisplayMinPay, OpeningBalance, BillLateCharge, EndingBalance, CreditLimit, TotalNetPayments, TotalNetPurchaseAndCash, TotalFeesBilled, Concessions, QuarterlyCreditScore and Bscore.

The model uses the bad variable as the response variable, which takes on binary values of 0 or 1, indicating whether a client is bad (1) or not (0). The model aims to predict the probability of a client going bad based on their predictor variables.

Tree Model

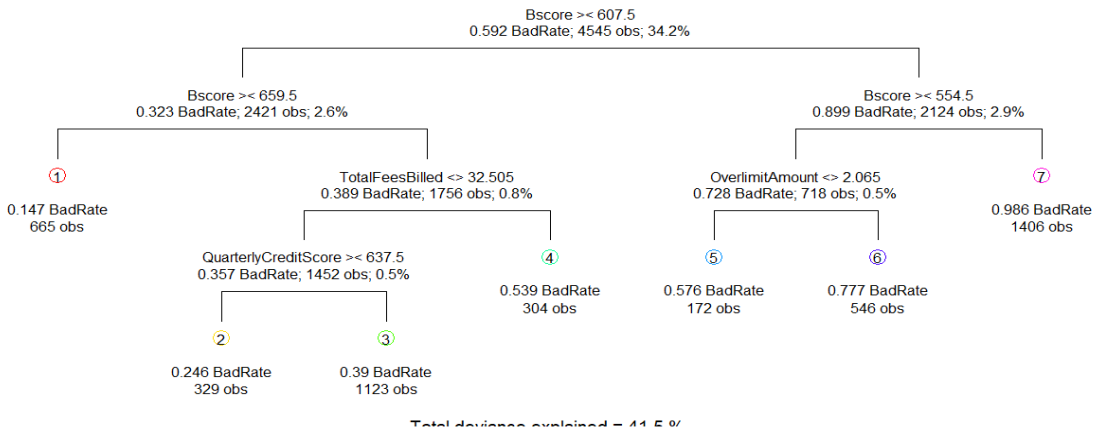

Figure 2.1: Tree model

After building the model in *figure 2.1* the following observations that were made:

- 665 customers with Bscore less than 659.5 a lower predicted probability of bad at 0.147.
- 329 customers with Bscore greater than 659.5 and Quarterly Credit Score less than 637.5 had a lower predicted probability of bad at 0.246.

- 564 customers with Bscore less than 554.5 and Over limit Amount less than 2.065 had a predicted probability of bad at 0.567.

- 329 customers with Bscore greater than 659.5 and Quarterly Credit Score less than 637.5 had a lower predicted probability of bad at 0.246.

- 172 customers with Bscore less than 554.5 and Over limit Amount greater than 2.065 had a higher predicted probability of bad at 0.777.

- 665 customers with Bscore greater than 554.5 a higher predicted probability of bad at 0.986.

2.1 Logistic Regression

Logistic regression is a statistical modeling technique that allows us to estimate the probability of an event occurring based on a set of predictor variables. In this study, we use logistic regression to model the probability of college graduation of enrolled students based on HSGPA and ACT score.

COEFFICIENTS OF THE LOGISTIC REGRESSION MODEL

The logistic regression model yielded the results below.

Table 5: Coefficients of the logistic regression model

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	1.11E+01	1.44E+00	7.692	1.45E-14	***
ExtStatusC	1.19E+00	5.16E-01	2.31	0.0209	*
ExtStatusE	**-1.45E-01**	5.20E-01	-0.279	0.7802	
ExtStatusF	-1.01E+00	7.34E-01	-1.379	0.168	
ExtStatusI	7.56E-01	1.30E+00	0.583	0.5596	
ExtStatusL	-8.35E+00	1.97E+02	-0.042	0.9662	
ExtStatusZ	6.85E-01	2.38E+00	0.287	0.774	
DaysDeliq	-9.15E-03	1.08E-02	-0.85	0.3953	
IntStatusD	-1.23E-01	3.60E-01	-0.342	0.7326	
IntStatusN	-5.00E-01	1.07E+00	-0.466	0.6415	
IntStatusO	8.41E-02	1.30E-01	0.648	0.5171	
IntStatusX	2.84E-01	2.98E-01	0.952	0.341	
ActualMinPay	3.41E-03	1.56E-02	0.219	0.8263	
OverlimitAmount	1.03E-02	1.61E-02	0.641	0.5217	

DisplayMinPay	-3.12E-03	1.54E-02	-0.202	0.8398	
OpeningBalance	1.38E-03	2.01E-03	0.685	0.4935	
BillLateCharge	-2.01E-03	7.00E-03	-0.287	0.7741	
EndingBalance	2.58E-04	2.08E-03	0.124	0.9014	
CreditLimit	-3.17E-03	5.33E-04	-5.945	2.76E-09	***
TotalNetPayments	1.91E-04	2.08E-03	0.092	0.9266	
TotalNetPurchaseAndCash	1.07E-03	2.08E-03	0.513	0.608	
TotalFeesBilled	2.51E-02	5.76E-03	4.357	1.32E-05	***
Concessions	2.73E-03	4.89E-03	0.558	0.5765	
QuarterlyCreditScore	-1.02E-03	3.21E-04	-3.176	0.0015	**
Bscore	-1.70E-02	2.24E-03	-7.589	3.22E-14	***

Based on the output, the coefficients that are statistically significant at the 5% level of significance are:

- Intercept (p-value < 0.001)
- ExtStatusC (p-value = 0.0209)
- CreditLimit (p-value < 0.001)
- TotalFeesBilled (p-value < 0.001)
- QuarterlyCreditScore (p-value = 0.0015)
- Bscore (p-value < 0.001)

The intercept estimate represents the expected log-odds of being in the "bad" category (i.e., having a defaulted credit account) when all the predictors are equal to zero. In this case, the intercept estimate is 11.09, which means that the odds of being in the "bad" category are exp(11.09) = 66,268 times higher than the odds of being in the "good" category when all other predictors are zero.

The ExtStatusC estimate is 1.191, which means that the log-odds of being in the "bad" category are 1.191 times higher for individuals with a status of "C" (compromised) compared to individuals with a current status of "A" (active). This estimate is statistically significant with a p-value of 0.0209.

The CreditLimit estimate is -0.003169, which means that a one-unit increase in CreditLimit (measured in thousands of dollars) is associated with a decrease in the log-odds of being in the "bad" category by -0.003169 units. This estimate is statistically significant with a very small p-value of 2.76e-09.

The TotalFeesBilled estimate is 0.02512, which means that a one-unit increase in TotalFeesBilled (measured in thousands of dollars) is associated with an increase in the log-odds of being in the "bad" category by 0.02512 units. This estimate is statistically significant with a very small p-value of 1.32e-05.

The QuarterlyCreditScore estimate is -0.00102, which means that a one-unit increase in QuarterlyCreditScore is associated with a decrease in the log-odds of being in the "bad" category by -0.00102 units. This estimate is statistically significant with a p-value of 0.0015.

The Bscore estimate is -0.01697, which means that a one-unit increase in Bscore is associated with a decrease in the log-odds of being in the "bad" category by -0.01697 units. This estimate is statistically significant with a very small p-value of 3.22e-14.

Section III: MODEL COMPARISON

3.1 Predicted Mean Response vs Actual

The plot of mean prediction and mean actual responses against depth in *figure 3.1* shows how well the model is performing in terms of predicting the response variable. The mean prediction line represents the expected response based on the model's predictions, while the mean actual response line represents the actual response in the data. If the mean prediction and mean actual response lines are close together, it indicates that the model is making accurate predictions. However, if there is a large difference between the two lines, it suggests that the model is not performing well and may need to be improved.

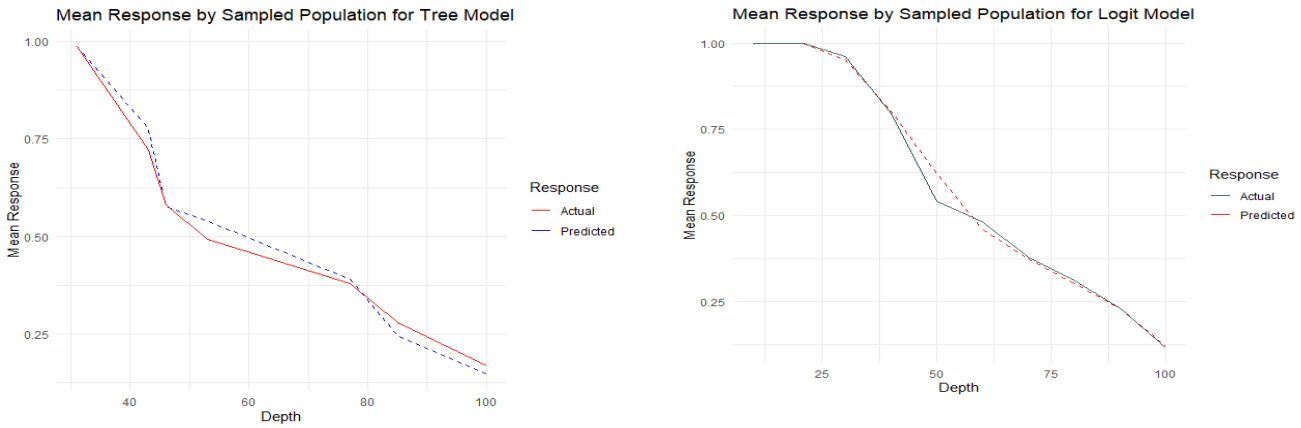

Figure 3.1: Actual vs. Predicted values

In the tree model, we observe that the mean prediction and mean actual response lines are closely aligned. This implies that the model may have a high accuracy in predicting the response variable.

Similarly, the logistic model shows almost perfect overlap between the predicted mean response and the actual mean response, indicating that the model can accurately capture the relationship between the predictor variable (bad) and the outcome variables without underfitting or overfitting the data.

The close overlap between the predicted and actual lines indicates that the model is a good fit for the data and can generalize well to new data. Moreover, the similarity in shape between the predicted and actual lines indicates that the model is making few, if any, major systematic errors in its predictions.

3.2 Gains Table

TREE GAINS TABLE

Table 6: Gains table for the Tree Model

Depth of File	N	Cume N	Mean Resp	Cume Mean Resp	Cume Pct of Total Resp	Lift Index	Cume Lift	Mean Model Score
31	933	933	0.99	0.99	52.2	170	170	0.99
43	355	1288	0.72	0.91	66.8	124	157	0.78
46	119	1407	0.58	0.89	70.7	100	152	0.58
53	185	1592	0.49	0.84	75.8	85	144	0.54
77	744	2336	0.38	0.69	91.8	65	119	0.39
85	244	2580	0.28	0.65	95.7	48	112	0.25
100	450	3030	0.17	0.58	100.0	29	100	0.15

The table suggests that as the depth of the model increases, the model's performance improves in terms of lift index and model score. The lift index indicates how much better the model is at predicting the target variable compared to random guessing. The model score is a measure of the model's accuracy.

At a depth of 31, the model has a high lift index of 170, indicating that it is performing significantly better than random guessing. The model score at this depth is also high, at 0.99. However, as the depth increases, the lift index and model score decrease. At a depth of 100, the lift index is 29 and the model score is 0.15.

LOGISTIC REGRESSION GAINS TABLE

Table 7: Gains table for the Logistic Regression Model

Depth of File	**N**	**Cume N**	**Mean Resp**	**Cume Mean Resp**	**Cume Pct of Total Resp**	**Lift Index**	**Cume Lift**	**Mean Model Score**
10	313	313	1	1	17.80%	172	172	1
21	309	622	1	1	35.30%	172	172	1
30	287	909	0.96	0.99	50.90%	165	170	0.95
40	303	1212	0.8	0.94	64.60%	137	162	0.8
50	303	1515	0.54	0.86	73.90%	93	148	0.62
60	303	1818	0.48	0.8	82.20%	83	137	0.46
70	304	2122	0.38	0.74	88.70%	65	127	0.37
80	302	2424	0.31	0.68	94.00%	53	118	0.3

90	303	2727	0.23	0.63	98.00%	40	109	0.23
100	303	3030	0.12	0.58	100.00%	20	100	0.12

For each depth level, the table shows the number of observations in the dataset (N), the cumulative number of observations up to that depth level, the response rate (Resp) which is the proportion of observations in the dataset that have a positive response, the mean of the responses up to that depth level (Mean), the cumulative percentage of observations up to that depth level (Cume Pct), the lift index which is the ratio of the response rate at that depth level to the overall response rate, the cumulative lift which is the cumulative ratio of response rate at that depth level to the overall response rate, and the model score which is the proportion of positive responses predicted by the model at that depth level.

From the table, we can see that the model performs best at depth 10 and 21 with a model score of 1, which means that 100% of the positive responses are correctly predicted by the model at that these depth levels. As the depth increases, the model score decreases and the percentage of observations with positive responses in the dataset decreases, indicating that the model becomes less accurate at predicting positive responses.

3.3 ROC

The plot in *figure 3.3* shows that the ROC curve for the Logistic Regression Model is consistently above the ROC curve for the Tree model. This indicates that the Logistic Model has better performance than the Tree model across all FPR values. Additionally, both the curves have similar shape, indicating that it has a similar ability to differentiate between true positives and false positives.

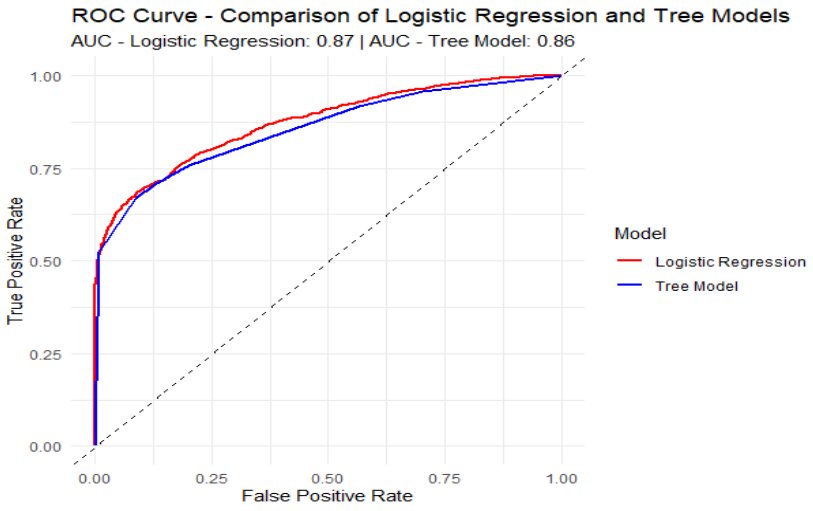

Figure 3.3: ROC curve

The AUC value measures the overall performance of the model, with a value of 1 indicating perfect classification and 0.5 indicating random classification. The AUC values for the two models suggest that the Logistic Model has a slightly better overall performance than the Tree model. The AUC value for the Tree Model is 0.86, which suggests that it performs better than a random classifier, whereas the AUC value for the Logistic Regression model is 0.87, which suggests that it also performs better than a random classifier.

Therefore, based on the plot and the AUC values, we can conclude that the Logistic Model outperforms the Tree model in this scenario. However, it is important to note that the choice of which model to use ultimately depends on the specific problem at hand and the trade-offs between model performance, interpretability, and other factors.

3.4 K-S

KS plot and statistic are commonly used to evaluate the accuracy of binary classification models, where the predicted and actual values are either 0 or 1. The KS statistic is a useful metric to compare the performance of different models and can help to identify which model is better suited for a particular task.

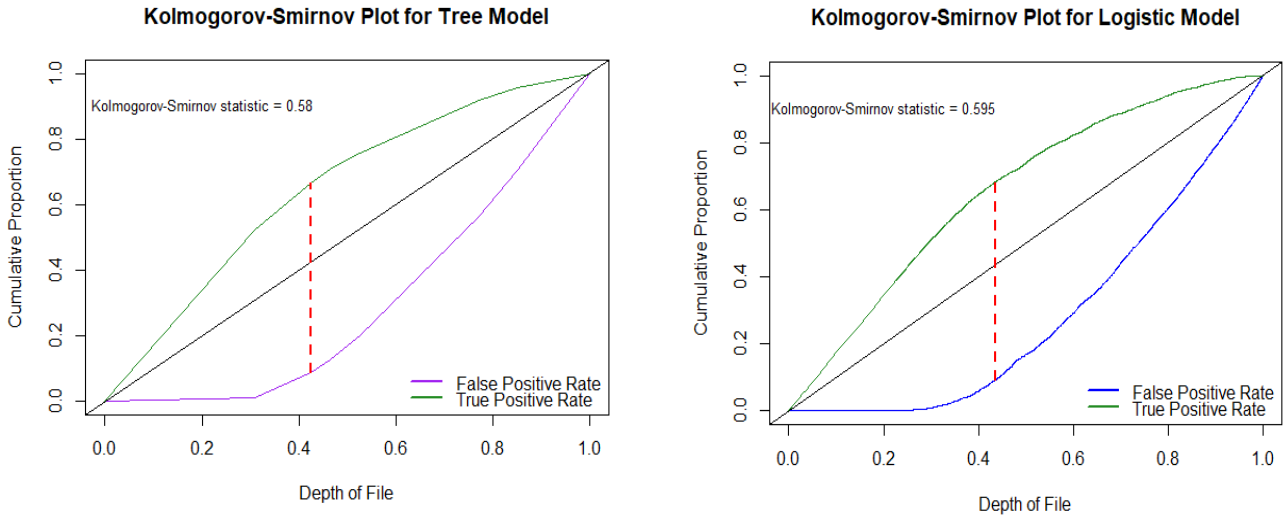

Figure 3.4: K-S plots

The KS plot generated for both models show a consistently higher TPR with a lower FPR, indicating a desirable outcome for a classification model. The KS statistic for the logistic model is 0.595, indicating slightly better performance in separating the positive and negative classes compared to the tree model, which has a KS statistic of 0.58. Therefore, the logistic regression model performs better in classifying the target variable.

Section IV: Findings and Conclusion

It can be concluded that the logistic regression model outperforms the decision tree model in predicting whether a customer will go bad or not. The ROC curve, gains table, AUC metric, and KS plot all indicate that the logistic regression model has better predictive power than the decision tree model.

Furthermore, the logistic regression model also provides insights into the most important features for predicting customer going bad. This information is valuable in understanding the factors that contribute to a customer going bad and can be used to develop strategies to prevent customers from going bad.

Therefore, it is recommended to use the logistic regression model over the decision tree model for predicting whether a customer will go bad. This will ensure that the predictions are more accurate and reliable, and the insights from the model can be used to develop effective strategies for managing customer risk.