A Comparison of Decision Tree with Logistic Regression Model for Predicting Behavioral Business Risk Score

Why Behavioral Scoring?

- Behavioral Scoring is used to model the usage and repayment behavior of consumer/businesses

- These models are used by lenders to adjust credit limits and to decide on the operational and market policy applied to each customer

- Behavioral scoring allows banks to rank customers from low risk to high risk in terms of their default likelihood

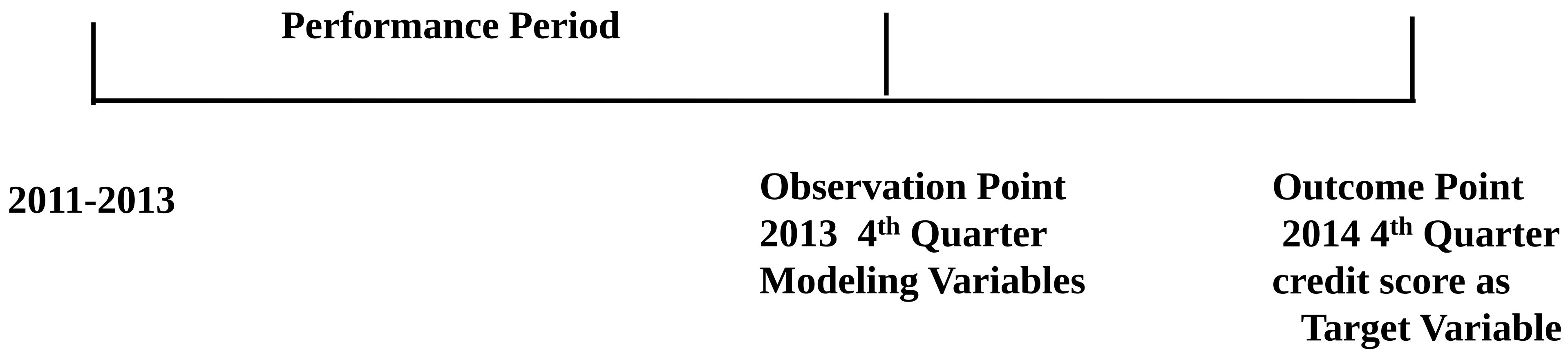
- Behavior scoring timeline

    - ➢ Observation Point
    - ➢ Performance Period
    - ➢ Outcome Point

- Observation Point: A particular point in the customer/business history is considered as observation point

- Performance Period : A period preceding observation point is called as performance period. This period is typically between 12-24 months

- Outcome Point :A time after the observation period to asses the good/bad credit status of the customer.



- Observation Point: Last Quarter of 2013
- Performance Period: Last Quarter of 2011-Last Quarter of 2013
- Outcome Point: Last Quarter of 2014 (Predict whether the business will maintain Good Credit (>450) or Bad Credit (< 450)

Why Varying Duration of Performance Period?

- Conventional behavioral models use 12-24 month performance period
- Open Banking gives opportunity to mine transaction data
- As transaction data is streaming data we many not get more than 12 month transactional history
- Combining behavioral data with transaction data may improve the credit risk models
- So it is necessary to generate behavioral models with less than 12 month duration of performance period
- Hence, we built behavioral models with 3 month, 12 month and 24 month snapshot data

Data Discovery

- Thirty-six datasets in total. Each dataset represents a quarterly report between 2006 and 2014; snapshots were taken annually in January, April, July and October.

- Each dataset has over 11 million observations representing unique businesses and 305 potential predictors representing businesses' general information that contain region, zip code etc , account activities and financial credit information such as business credit risk score etc.

- For this project, considered the dataset October 2014

- The response variable is the business risk score of the companies.,Assumed a risk score below 450 can be considered "bad" for a company.

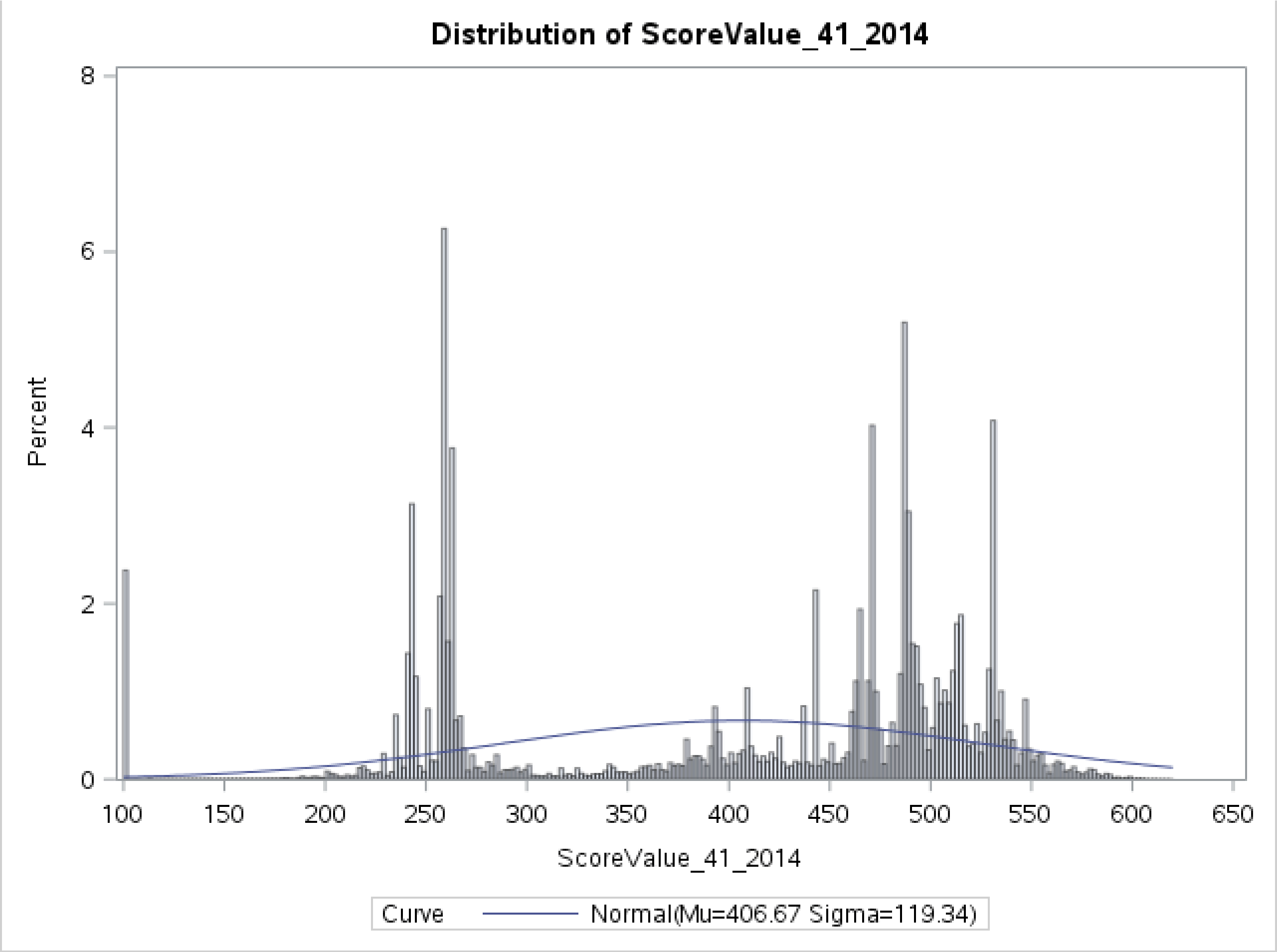


Table 1: Distribution of the Binary Dependent Variable BADCredit

The FREQ Procedure

BADCredit	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	2506813	53.41	2506813	53.41
1	2186558	46.59	4693371	100.00

Data Preprocessing

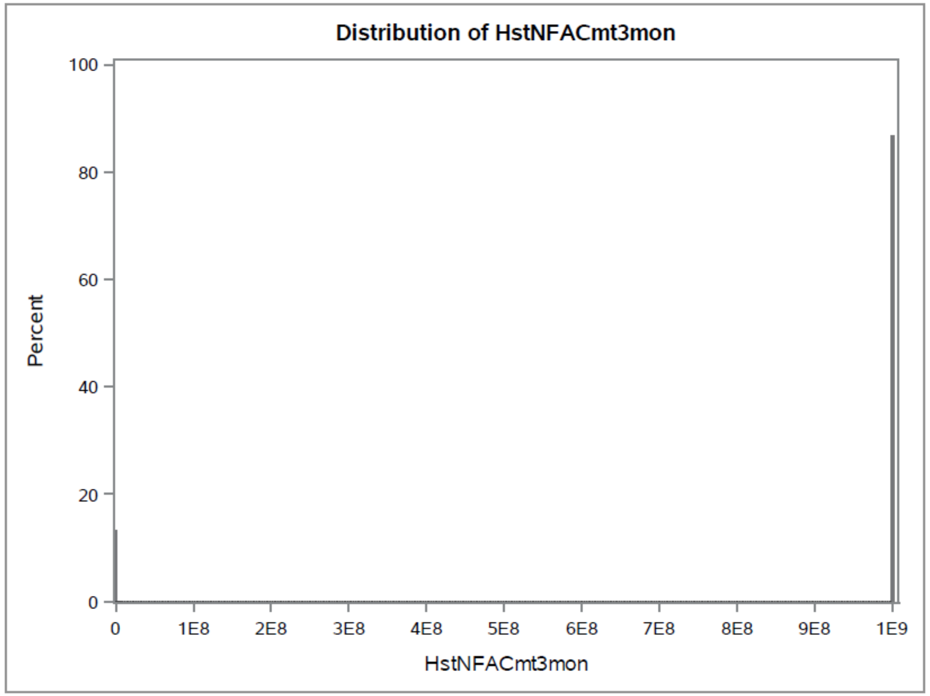
Dimensionality Reduction by Removing Variables with High ratio of Coded or Missing values:

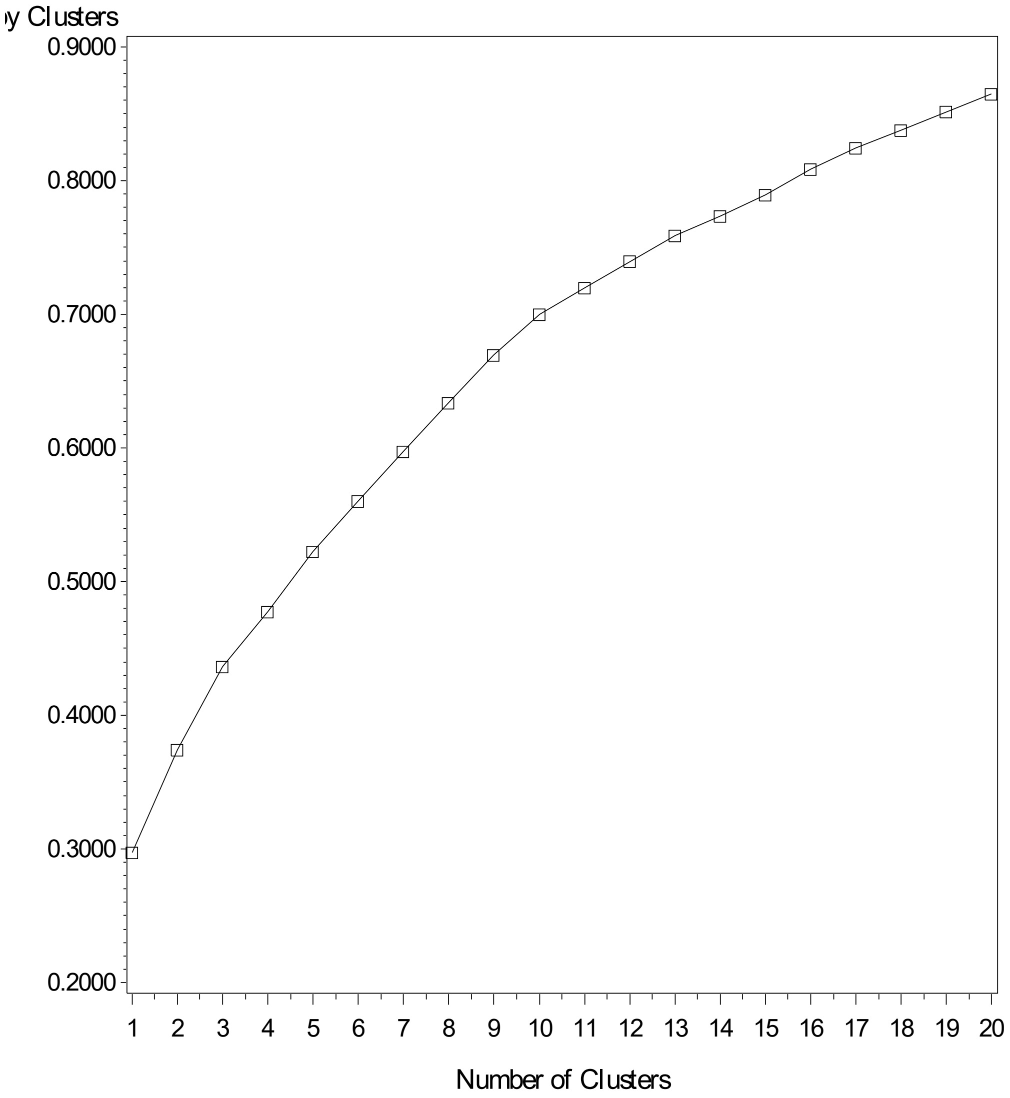
- Variable HstNFACmt3mon(Highest Non-Financial Account Limit Reported in Last 3 Months) has over 80% of the coded values.

- predictors with a high ratio (>50%) of coded or missing values were removed from the dataset.

- For some predictors where coded or missing data is less than 50%, median imputation strategy was used since most predictors are right skewed.

- Based on this criterion, we left with 67 variables (56 numeric and 11 categorical) .

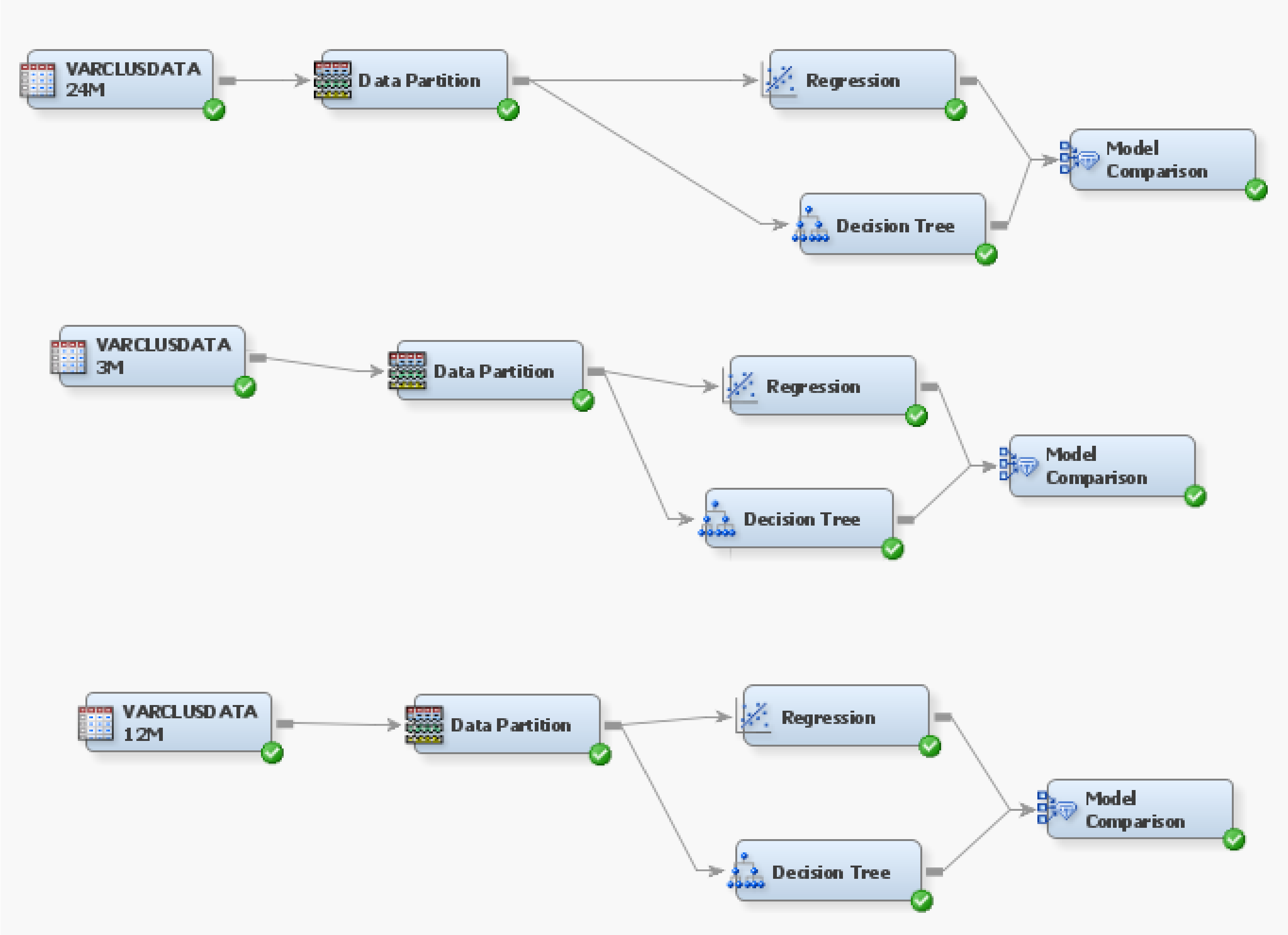




Dimensionality Reduction by Variable Clustering:

- Variable Clustering is an unsupervised technique to reduce variables by eliminating correlated variables
- Twenty clusters were created from 67 variables and these clusters explained around 85% of the variation in the data
- One variable is selected from each cluster which has highest correlation among its cluster members and lowest correlation with rest of the clusters
-

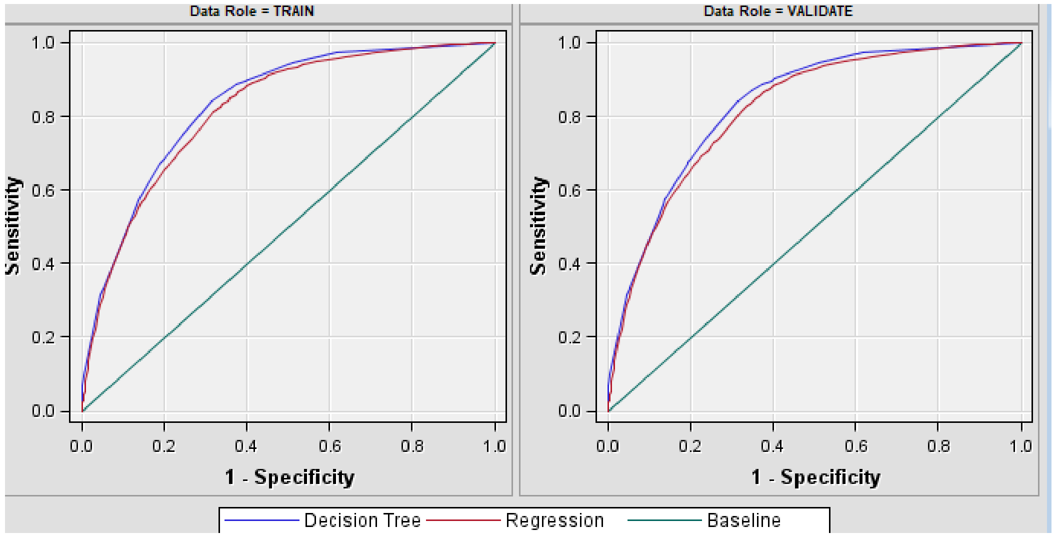
Data Modeling Architecture



Model	Accuracy	Type 1 Error	Type II Error	ROC
Logistic Regression	0.75	0.261	0.274	0.83
Decision Tree	0.85	0.108	0.201	0.91

Table 1: Performance of Models with 24 M variables

- Decision tree approach has a very good performance on the given data
- False positive rate and false negative rate are both lower in decision tree than in logistic regression.
- Moreover, the Decision tree model is parsimonious, it uses six lesser variables than logistic regression



Decision Tree Variable Importance

Variable Name	Importance
Number of Non-Financial Accounts Reported in the Last 12 Months	1.0000
Percent of Non-Financial Charge-Off Accounts to Total Accounts Reported in Last 24 Months	0.8915
Total Cycle 4+ Non-Financial Past Due Amount in Last 24 Months	0.4535
Year Started	0.3588
Large Business Indicator	0.3492
Lien / Judgment Indicator	0.3046
Total Non-Financial Past Due Amount in Last 24 Months	0.2696
Total Cycle 3 Non-Financial Past Due Amount in Last 24 Months	0.1732
Number of Employee Range	0.1525
Percent of Non-Financial Charge-Off Accounts to Total Accounts Reported in Last 3 Months	0.1511
MasterState	0.1435
Region	0.0728
Total Liabilities on All Liens	0.0580
GovernmentEntityFlag	0.0400
Total Liabilities on All Judgments	0.0243

Logistic regression Variable Importance

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
GovernmentEntityFlag	2	10613.6158	<.0001
Industry	9	773.0825	<.0001
LargeBusinessInd	2	2801.9765	<.0001
LienJudInd	1	46896.6223	<.0001
MasterState	52	15265.2863	<.0001
NoEmployeeRange	1	94.8579	<.0001
PublicCompanyFlag	2	22.6824	<.0001
SubsidiaryInd	2	21.8295	<.0001
TotUtiNFA	1	384.9172	<.0001
UltParentEntity	2	136.5002	<.0001
UltParentPublic	2	93.8113	<.0001
YearStarted	89	145972.965	<.0001
pctNFChgAccAcc24mon	1	603228.768	<.0001
totC3NFPDAmt24mon	1	699.6905	<.0001
totC4NFPDAmt24mon	1	7628.5367	<.0001
totLAllLiens	1	405.9725	<.0001
totNFA1CPD24mon	1	480.3924	<.0001
totNFA2CPD24mon	1	12183.6931	<.0001
totNFA3CPD24mon	1	3768.1277	<.0001
totNFA3CPDC24mon	1	91599.1739	<.0001
totNFPDAmt24mon	1	319.5851	<.0001

Business risk models with 12 & 3 month data

Model	Accuracy	Type 1 Error	Type II Error	ROC
Logistic Regression	0.75	0.231	0.263	0.84
Decision Tree	0.84	0.094	0.237	0.88

Table 2: Performance of Models with just 12 Month variables

Model	Accuracy	Type 1 Error	Type II Error	ROC
Logistic Regression	0.74	0.233	0.301	0.81
Decision Tree	0.83	0.118	0.227	0.87

Table 3: Performance of Models with just 3 Month variables

Summary

➢Off the initial 300 variables less than 15 variables are sufficient to predict the outcome (“Bad Credit’) after one year based on the two year performance period

➢ “Number of Non-Financial Accounts Reported in the Last 12 Months” , “Percent of Non-Financial Charge-Off Accounts to Total Accounts Reported in Last 24 Months” etc are important predictors

➢ Decision Tree model is not only parsimonious but also performed better than Logistic Regression model

➢ Scoring models with 3 month performance period has similar results as model with 24 month performance period