## Predicting the Loan Delinquency Category

The primary objective of this analytical exercise if to predict the type of delinquency category.

## Why to Identify Delinquent Loans - Business use case

Identifying delinquency category of loans is only one piece of the big puzzle. The bigger picture is to identify the default rate, prepay rate and loss severity vectors. The predictive models generate future cashflow and vectors for different economic scenarios based on the individual business needs. Investors take the vectors/coefficients into considerations to evaluate bond’s future returns for investors

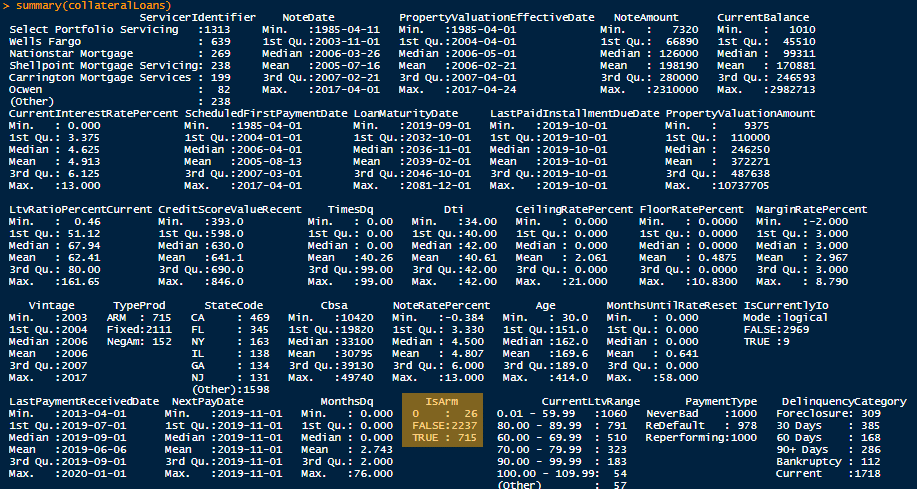
Following are the categories of the class variable Delinquency Category

|  |  |
| --- | --- |
| Category | Description |
| 30 days | Loan is 30 days delinquent. A loan can transition to either 60 days, 90 days or Current status again |
| 60 days | Loan is 60 days delinquent. A loan can transition to either 90 days or Current status again |
| 90 Days | Loan is 90+ days delinquent. A loan can transition to Current or Bankruptcy status |
| Bankruptcy | Loan is in Bankruptcy status. The loan can transition to Foreclosure or REO status |
| Current | If the loan is paid on time and is in no delinquent situation |
| Foreclosure | Post-Bankruptcy the loan is in foreclosure and is not considered to calculate future cash flows and predictive vectors (prepay, default and loss severity) |
| REO | Post-Bankruptcy the loan is in Real-estate owned status and is not considered to calculate future cash flows and predictive vectors (prepay, default and loss severity) |
| *For Modeling purposes, it makes sense to combine the status of Foreclosure and REO. Post- bankruptcy, the lender decides either to foreclose or own the property for fix and sale. We have merged the REO delinquency category to Foreclosure for modelling purposes* | |

## Summary of Data

The data we used is licensed from third party vendor that provides collateral/loan level data on a monthly basis for all securitized assets. We have used around 3000 loans from several RMBS deals (bonds) of various vintages and prospectus groups.

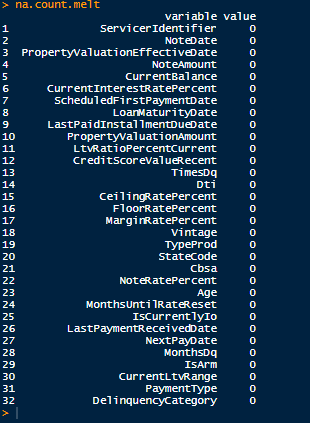
|  |  |
| --- | --- |
| Total records used to train and test | 2978 |
| Number of Columns | 33 |
| Number of Categorical Columns | 7 |



## Data Exploration

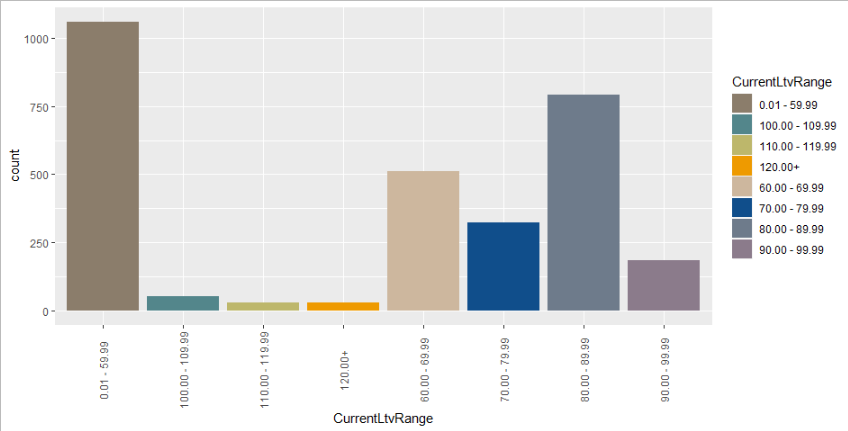
#### Missing Values

Since this is a post-transformation data, there are no missing values in all the key attributes that we have used. Only issue that we fixed is *IsArm* variable that has both 0 and false values. All the values are converted to 1 for true and 0 for false part of the cleansing routine



#### Data Coverage by Loan-to-Value

The graph below shows coverage the data by LTV (Loan-to-Value) range. In the mortgage industry, LTV below 80 is favored. Any loans request above the 80 LTV could potentially be charged higher interest rate. During the data transformation process, any missing values for LTV is given a default value of 80



#### Identifying Outliers for Key Numerical Predictor variables

As per the domain, Credit Score, Current Balance, Property Valuation Amount and Original Note Amount are key numerical factors to evaluate the delinquency status on the loan. The Current Balance and Property Valuation Amount are calculated every period for the updated delinquency factor evaluation. Current balance is retrieved from a third party vendor but the property valuation amount is calculated during the transformation process.

|  |
| --- |
|  |
|  |
|  |

#### Normalizing the Data

Based on the visualization shown below, Note Amount, Property Valuation Amount and Current Balance are right skewed. We did log() transformation to normalize the data by shortening the range of the values.

|  |  |
| --- | --- |
| *Note Amount - Prior to transformation* | *Note Amount - Post transformation* |
| *Current Balance - Prior to transformation* | *Current Balance – Post transformation* |
| *Property Valuation Amount - Prior to transformation* | *Property Valuation Amount – Post transformation* |

## Classification Models

We tried 4 different classification models to predict Delinquency Category. Initially, we used all the variables but as we tune our parameters with different subset combinations, number of tress etc., we kept the log of Misclassification Error rate and ultimately selected the model that reduced the misclassification error rate.

We used 75% of the data for training the model and the rest 25% for the validation set.

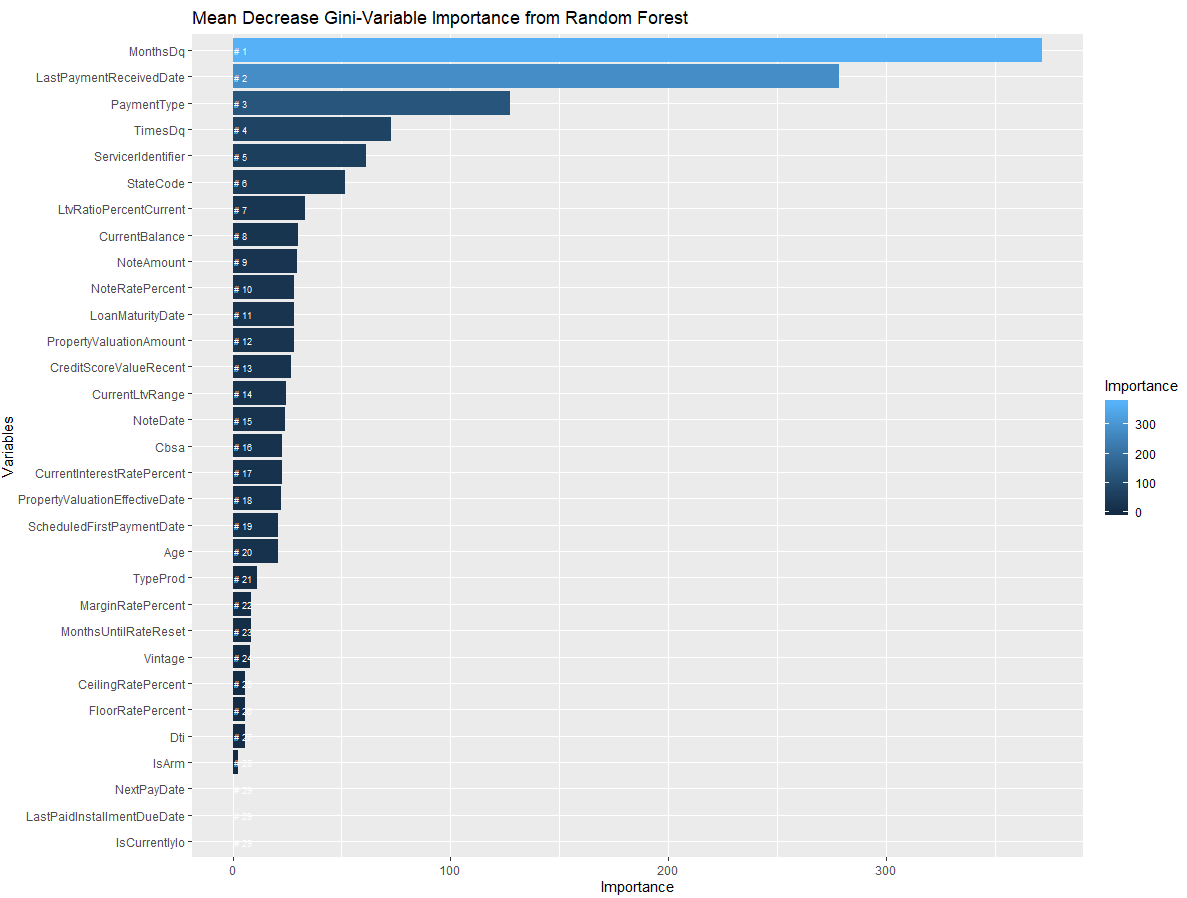
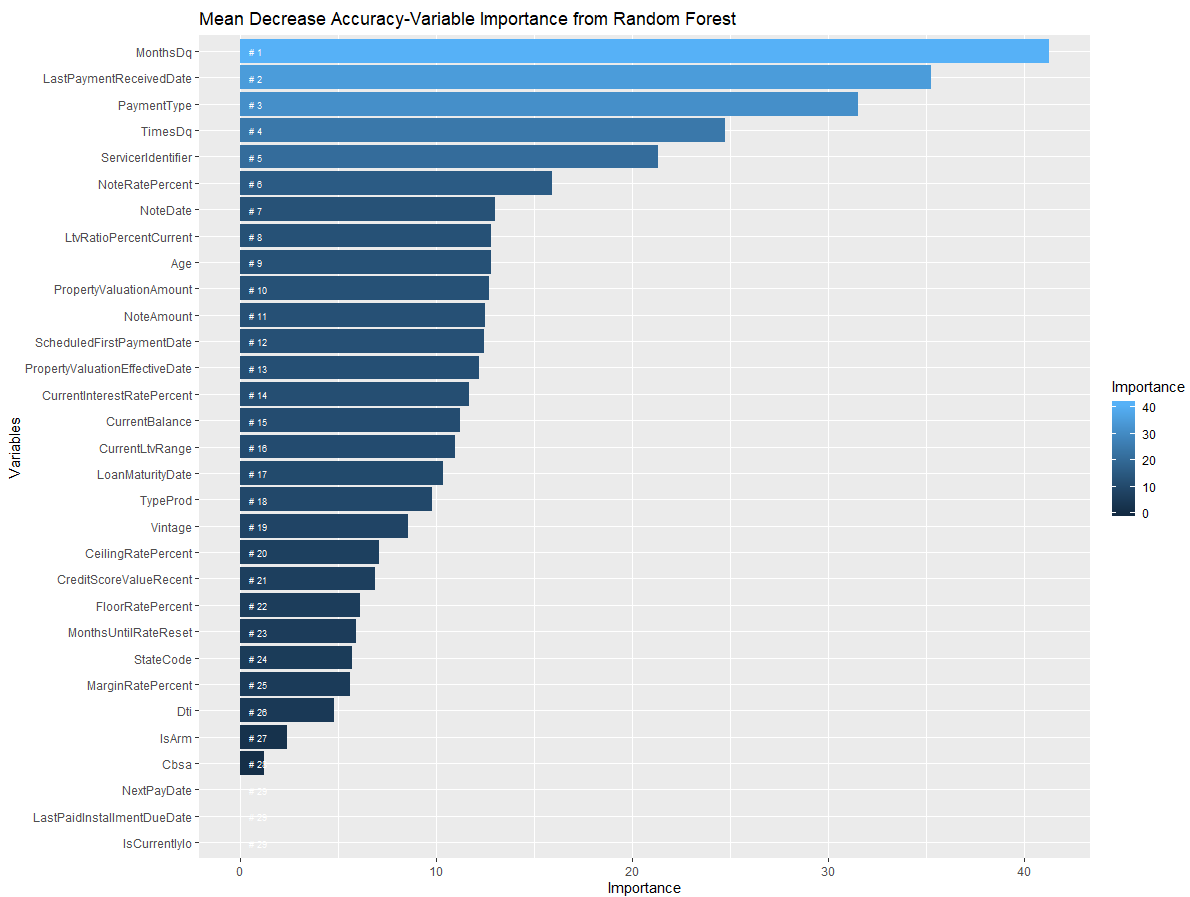
### Random Forest Model

Random forest was the first model that we ran and initially used all the variables to identify the important variables to consider for our future models. We used at least 6 difference combinations of target variables based on the Mean decrease Gini-variable and Mean decrease accuracy-variable importance as shown in the section below. We identified about 14 variables that provided the lowest MSE across all the models.

On Random forest specifically, we used different combinations of individual trees and number of randomly chosen variables at each split. In the end 1000 individual trees with 4 randomly chosen variable gave the lowest MSE.







#### Results of Random Forest Model

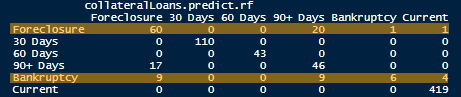
##### Accuracy and MSE

We have listed few results of difference combinations that we implemented.

|  |  |  |
| --- | --- | --- |
| Predictors | Accuracy | MSE |
| All | 90.20% | 9.798% |
| 14 variables mtry=5, ntrees=1000 | 91.54% | 8.45% |
| 14 variables mtry=4, ntrees=5000 | 91.81% | 8.185% |
| 15 variables mtry=4, ntrees=5000 | 91.40% | 8.59% |
| 14 variables mtry=4, ntrees=1000 | 91.82% | 8.18% |

##### Confusion Matrix

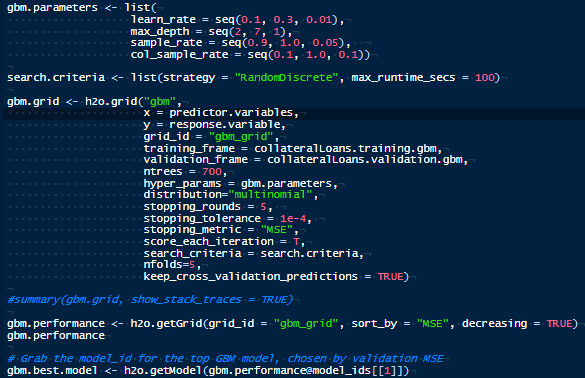
As we can identify that most of the Random Forest’s accuracy was lost due to Foreclosure and Bankruptcy classification error. When a debtor gets delinquent, he/she can either is forced to foreclose (initiated by lender) or file for bankruptcy (initiated by debtor). For modelling purposes, we could have easily combined the Foreclosure and Bankruptcy class to increase the accuracy of the model. But in reality there are some rules around DTI and Credit Score of the debtor that could be utilized to distinguish between these two classes. Class variables Current, 30 days and 60 days are predicted perfectly. I was a little surprised that 4 Bankruptcy loans were falsely labelled as Current.



### GBM (Gradient Boosting Machines) – using H2o Package

We used the H2o package for running GBM model as it could get intensive capacity wise on a single machine. H2o is a Java Virtual machine that is optimized for doing in memory processing of distributed, parallel machine learning algorithms on clusters. A cluster is a software construct that can be can be fired up even on our laptops.

##### Parameters for GBM

We tried multiple combinations of parameters before finalizing the parameters that gave us the least MSE. We used Cross Validation to identify the best suited model using 5 folds.

Following are the major parameters that were used in the final model

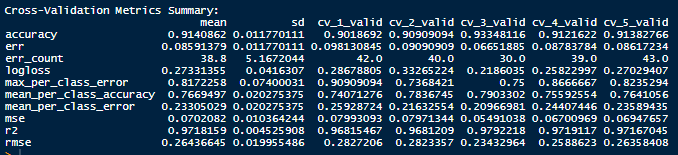
|  |  |
| --- | --- |
| Number of trees | 700 |
| List strategy | RandomDiscreet |
| Hyper Parameters | Learn\_rate = seq(0.1,0.3,0.01)  Max\_depth = seq(2,7,1)  Sample\_rate = seq(0.9, 1.0, 0.05)  Col\_sample\_rate = seq(0.1, 1.0, 0.1) |
| Distribution | Multinomial |
| Stopping metric | MSE |
| Number of Folds for Cross validation | 5 |

##### Best Model Used – Post Cross Validation

Our stopping metric for GBM was MSE. We sorted the model based on the lowest MSE and picked the top model to predict our validation set. The top model has a MSE of .1045 compared to the highest MSE of 0.548

##### 

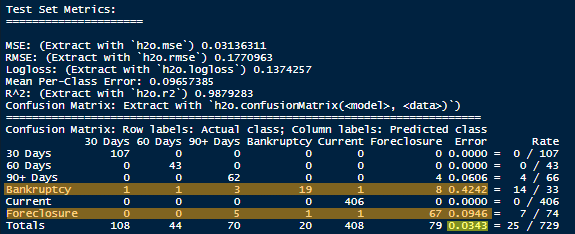
Below are few statistics on the best model that we will be using



##### Confusion Matrix/ Accuracy and MSE

We expected, and rightly so, for GBM to have a better accuracy rate compared to Random Forest just for the fact that it is so computational intense and we used cross validation to identify the best model to use. Our fear is that due to cross validation the model might have gotten highly complex and variant. Compared to Random Forest, GBM had a much lower MSE of **3.43%** with accuracy of **96.57%**

The interesting thing here is that GBM did very well classifying the Foreclosure, Bankruptcy and 90+ classes compared to Random Forest model. Both models did very well on Current class variable as I believe it was straight forward. When the predictor MonthsDq is 0, it always mean that loan is in good current status.



### Logistic Regression

We have used multinomial logistic regression since our outcome variable contains multiple classes and are nominal.

### Stepwise Logistic Regression – Forward

In this approach, the number of predictor variables is reduced automatically for the model with better accuracy.