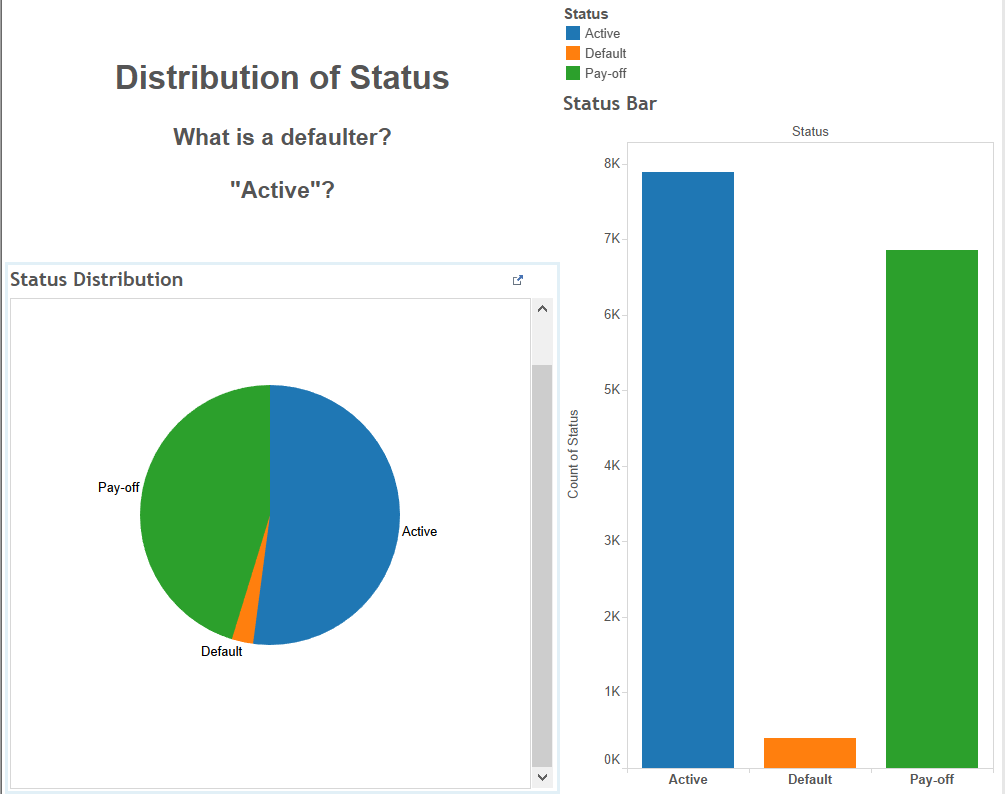
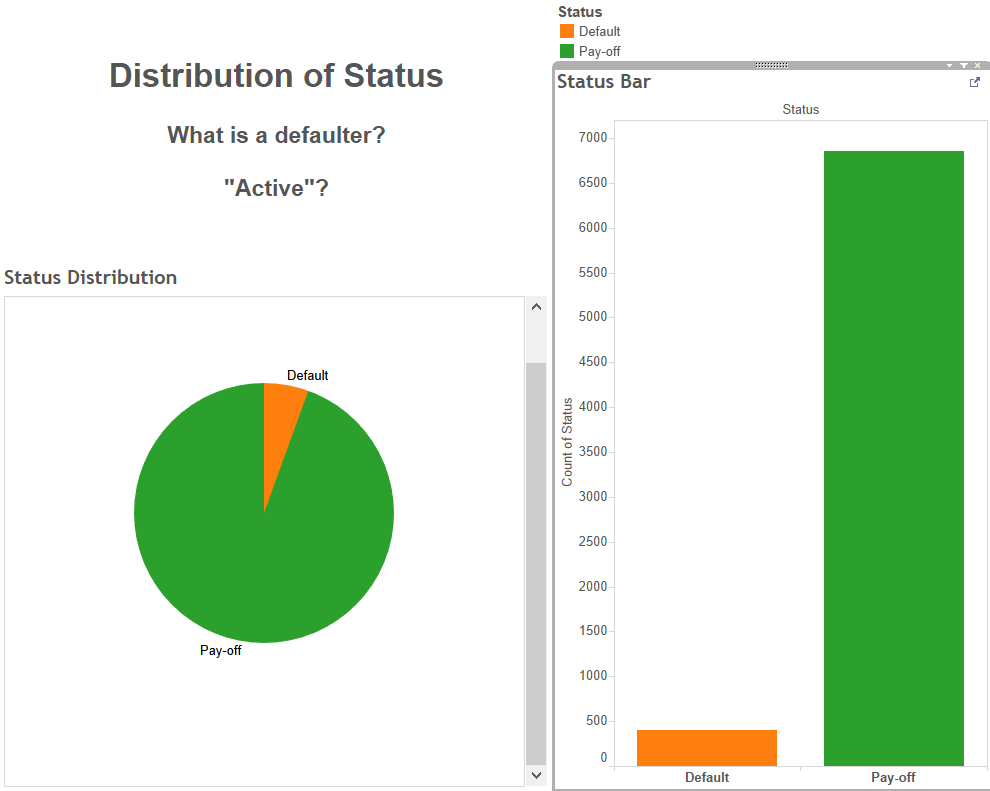
**2 Mortgage Defaults**

2.1 Exploratory data analysis

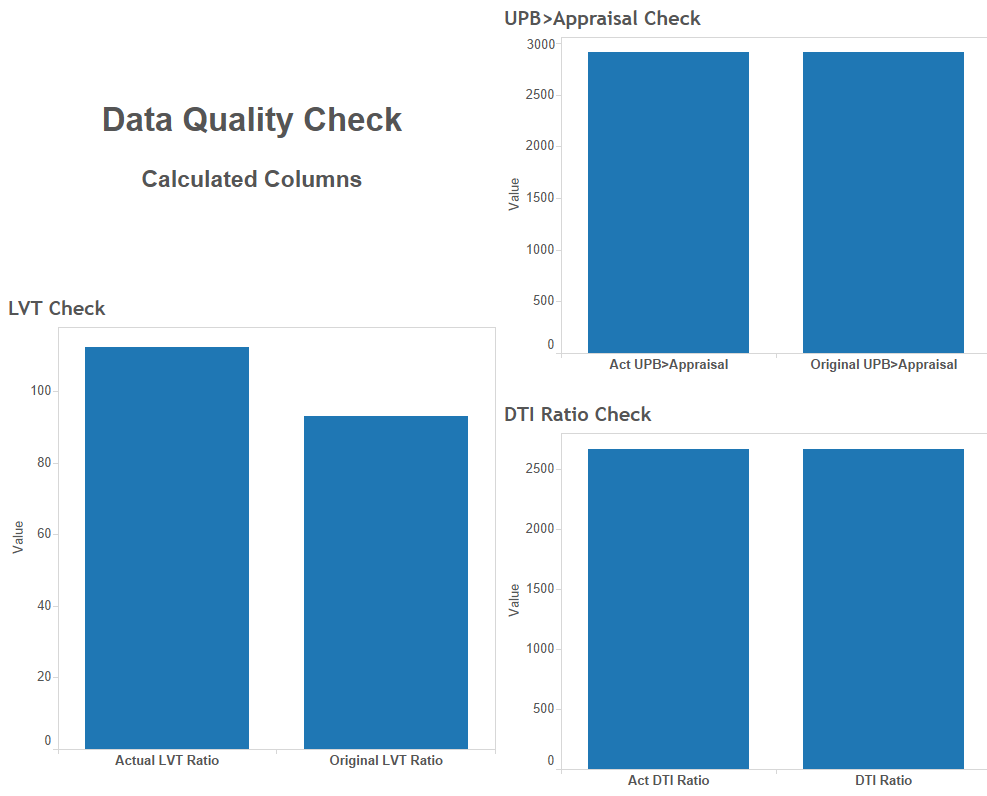
Tool: Tableau.

Steps:

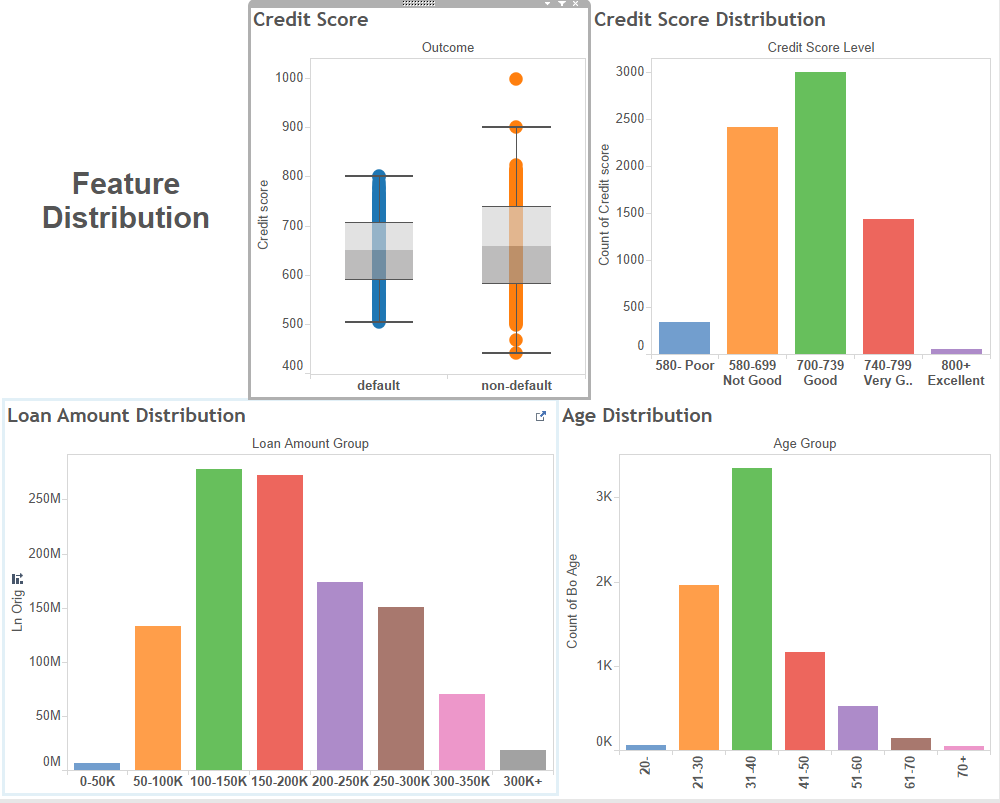
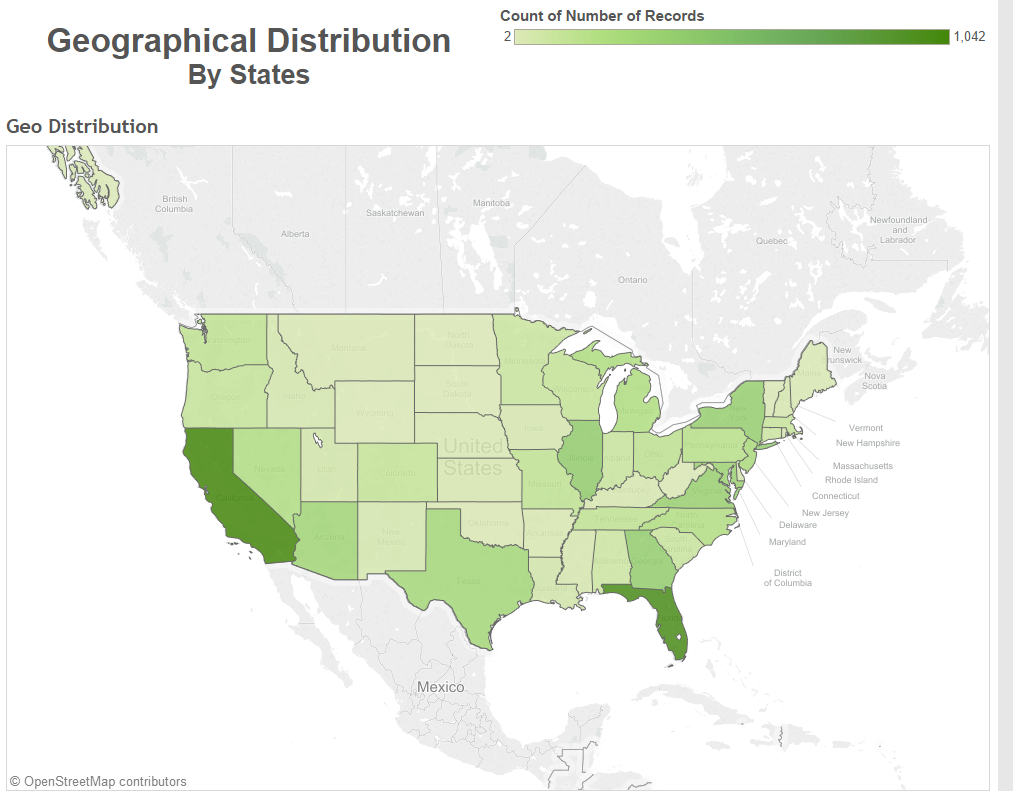
1. According to the chart, noticed that the group of mortgage status “Active” has been classified as non-default. However, in the real life, these active mortgage may turn into a default as soon as the borrowers are no longer able to pay for that. Therefore, all the “active” records should be dropped from the non-default group, which decrease the record number by 7893.

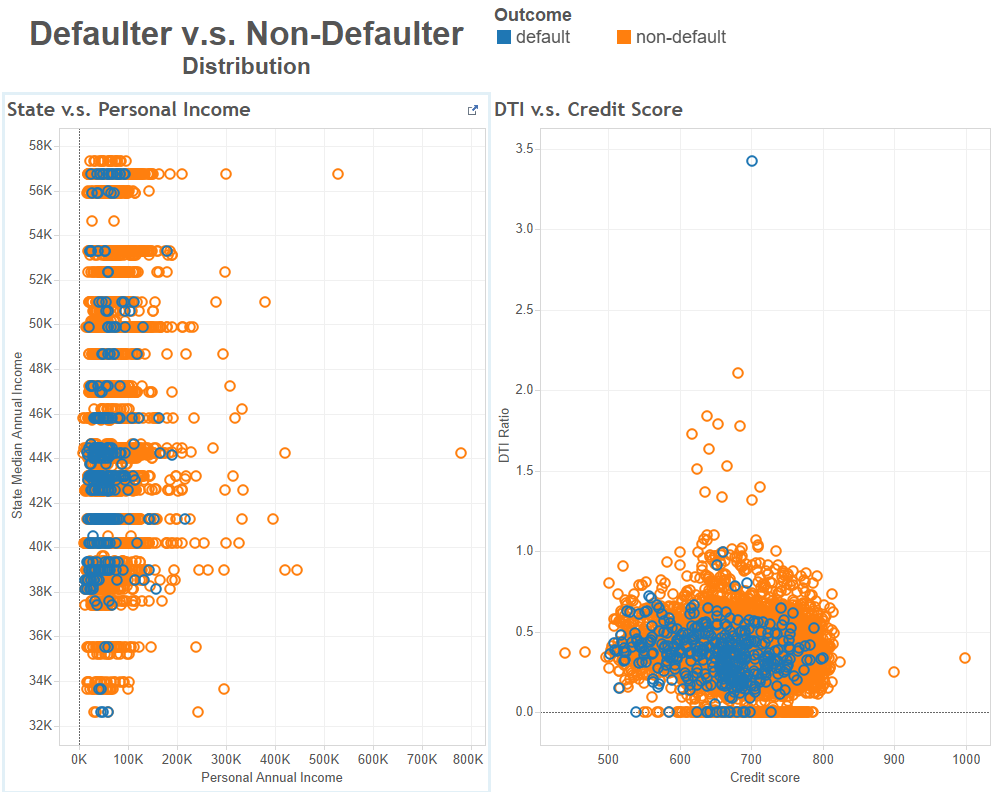
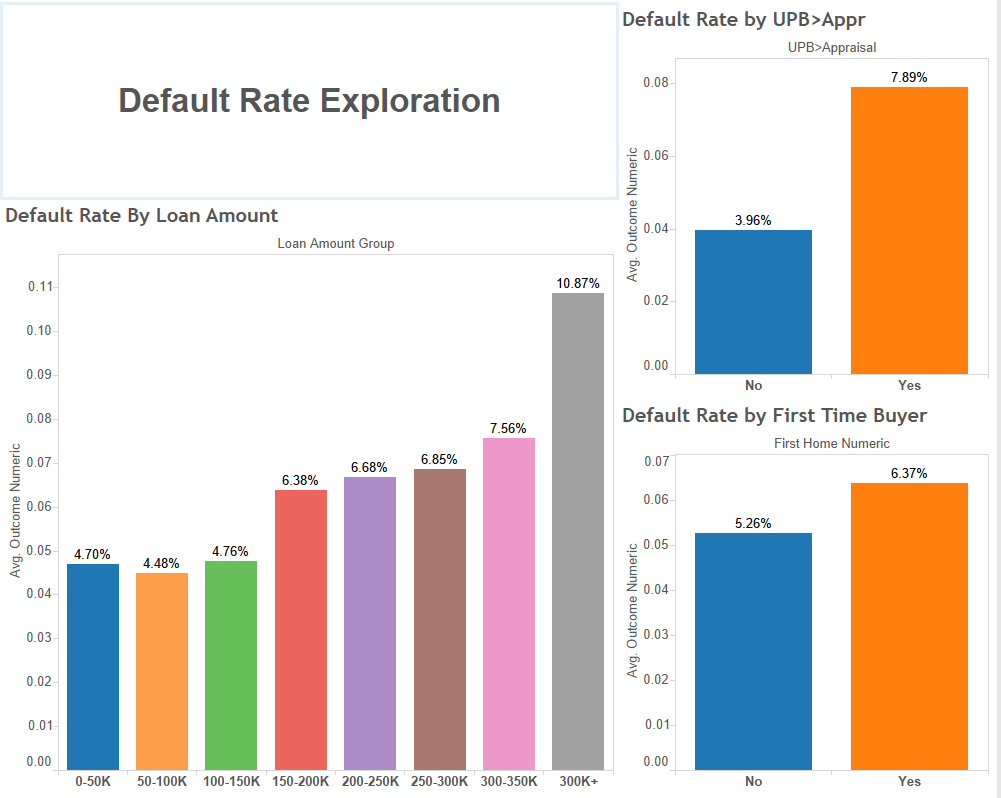
 

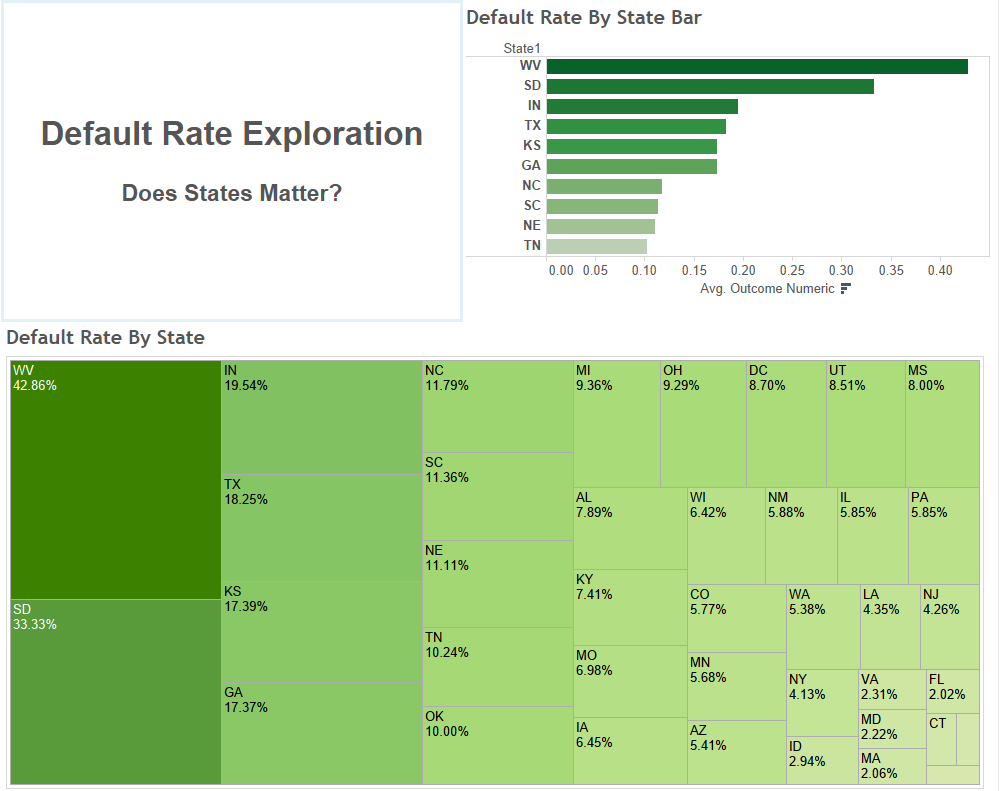
1. There are calculated columns in the dataset need to be checked by creating new calculated columns in Tableau. After the process, the data inaccuracy on LVT ratio have been discovered. The new calculated column will be used as the correct data for further analysis.



1. Further data exploration has been performed in multiple ways. First, mortgage distribution using different features has been visualized using bar chart, geo-map and box plot. Second, in order to explore the relationships between default/non-default and different features, scatter plot charts have been used. Third, because of the significant sample size difference between default and non-default, the KPI default rate has been created to measure the impact of different features. Bar charts and heat maps are used in default rate analysis.



2.2 Build classification models

Tool: XLMiner

Steps:

1. Data screening and pre-processing:
   1. Filter out the rows having “active” in Status column.
   2. Create calculated column “Act\_LTV\_Ratio” to replace the “Orig\_LTV\_Ratio\_Pct”. The formula used is Ln\_Orig divided by the smaller between orig\_apprd\_val\_amt and pur\_prc\_amt. If orig\_apprd\_val\_amt = 0, then use pur\_prc\_amt[[1]](#footnote-0).
   3. Drop unused column: Status, state.
   4. Convert binary categorical columns’ values into 1 and 0: First\_home, OUTCOME.
2. Outlier identification and data cleansing: according to the result of visualization. The following outliers have been identified and need to be deleted from the dataset.
   1. One column has a credit score of 999, and with monthly income of $1,187, the borrower managed to pay-off a $228,000 loan. The house price is $50,000, which is far below the loan amount. This row shows obvious abnormal features of fraud data, and need to be deleted from the original dataset.
3. Partition data
   1. Data partition has been performed through standard partition option with a 60% training and 40% testing. Random seeds has been set as 12345. Same training and testing partition will be used in all the classification models.
4. Build models
   1. Set cut-off values. Cut-off value has been set to 0.05538 (5.538%), the generic percentage of default for the dataset.

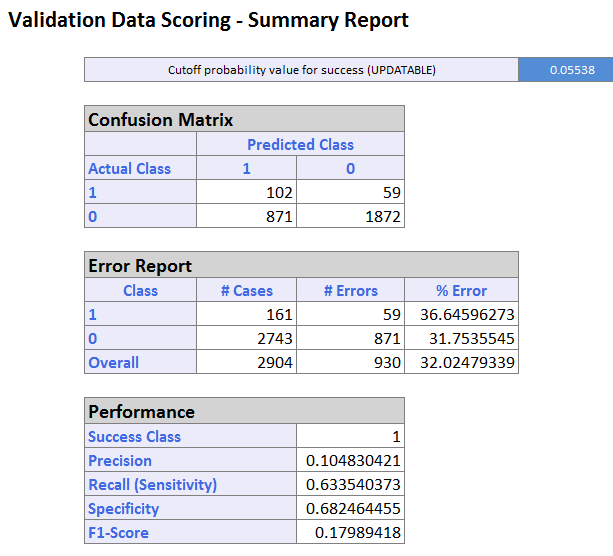
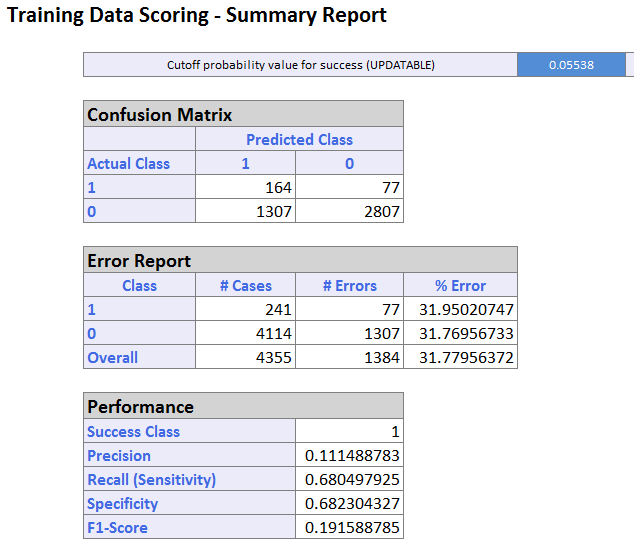
Reason: According to the scenario, the bank (lender)’s goal is to identify the potential defaulters so that they can purchase secondary insurance to prevent potential loss. Also, the cost of misidentifying a non-defaulter as a defaulter is much less than missing a real defaulter. Therefore, the model is expecting an unfair measurement of accuracy for a two-side classification. A cut-off value of 50% is not applicable in this case.

* 1. Based on the cut-off value. Classification has been performed using the following machine learning algorithms: Logistic regression, CART (Classification and Regression Tree) and Random forest. Default settings are applied for CART & Random Forest.

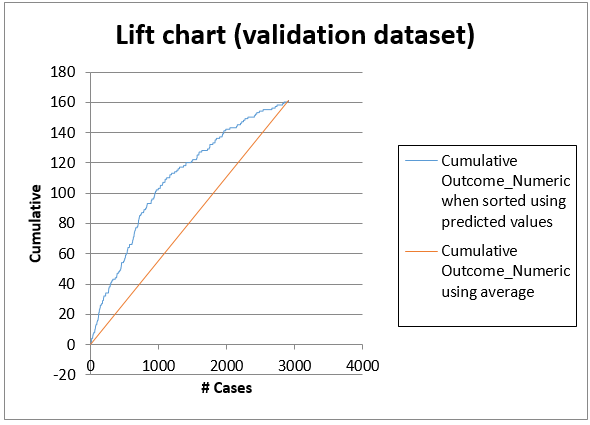
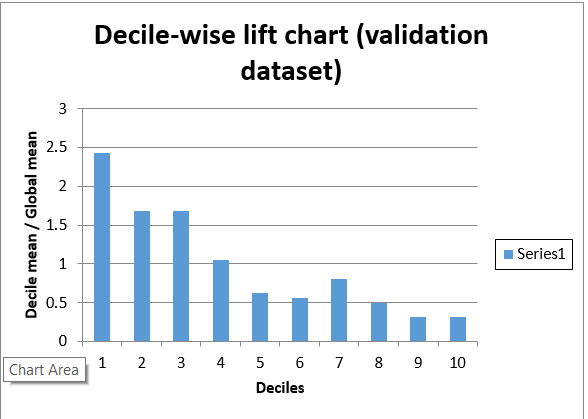
2.3 Performance evaluation

Measurement: confusion matrix. Lifting chart. Decile-wise lift chart.

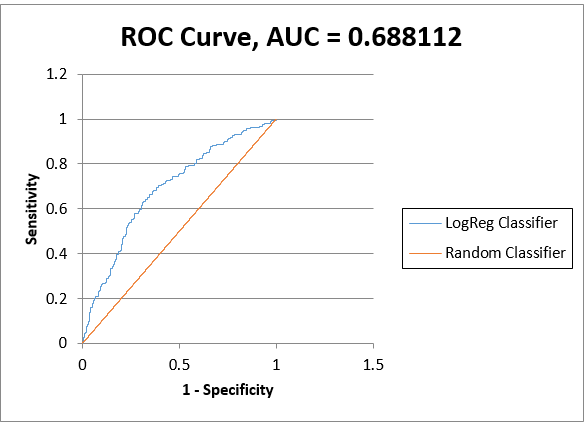
1. Logistic Regression:
   1. In the confusion matrix of training dataset, the error rate for defaulter and non-defaulter are both close to 32%, while in the testing dataset, the error rate of defaulter increased to 36.6% and for non-defaulter it still remain lower than 32%. The overall error rate for both training and testing are similar (32%)



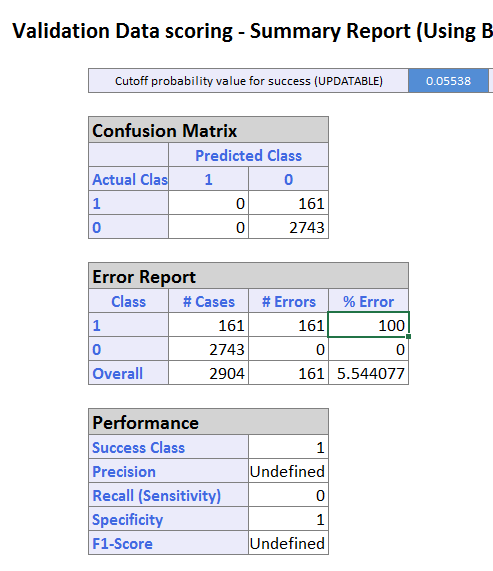
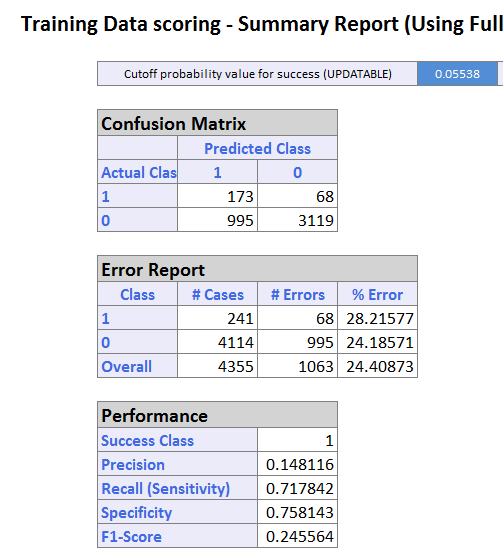
* 1. In the lift chart for testing dataset, the curve is above the straight line which indicates random classification rules. The Decil-wise lift chart also has higher head and lower tail. Both lift chart indicate the classification model performs better than a random model.

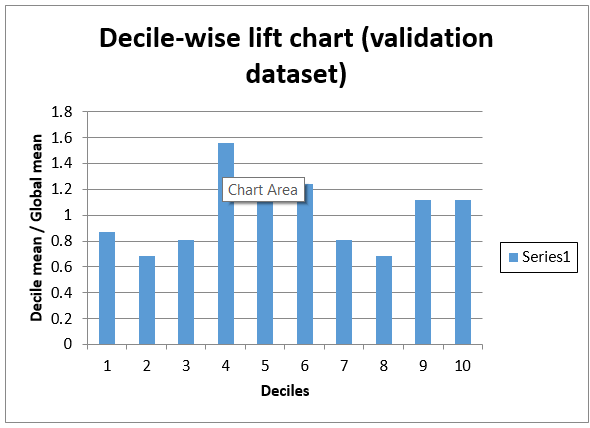
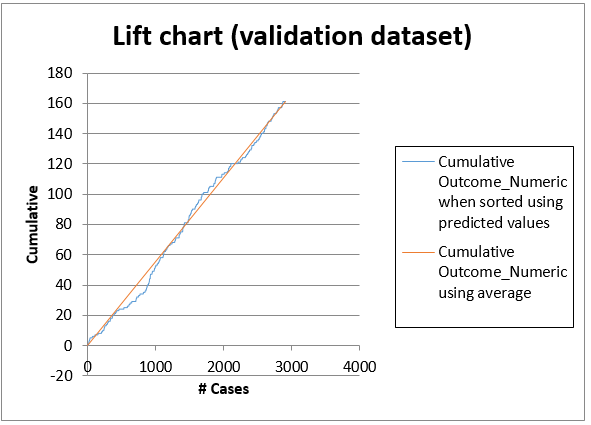
* 1. Roc Curve for testing dataset is above the straight line which indicates random classification rules. AUC = 0.688112, which also indicate the logReg Classifier performs better than a random classifier.



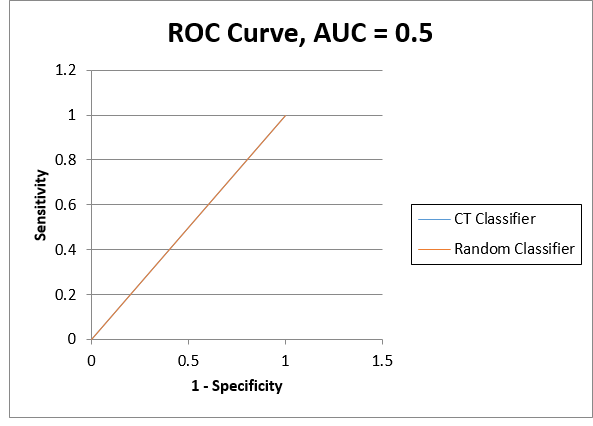
1. CART
   1. In the confusion matrix of training dataset, the error rate for defaulter is 28% and for non-defaulter is 24%, while in the testing dataset, the error rate of defaulter increased to 100% and for non-defaulter dropped down to 0%. The overall error rate for training is 24% and for testing is 5%



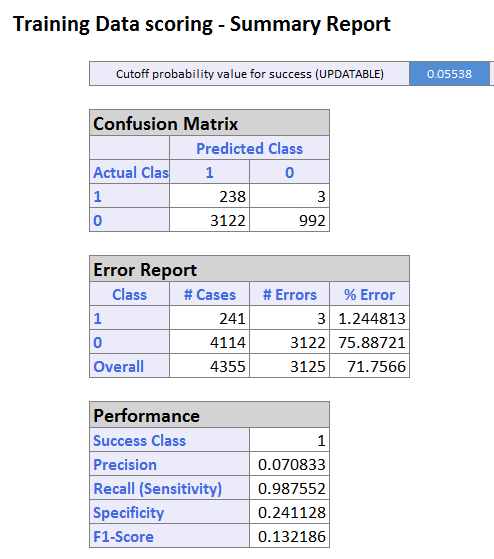
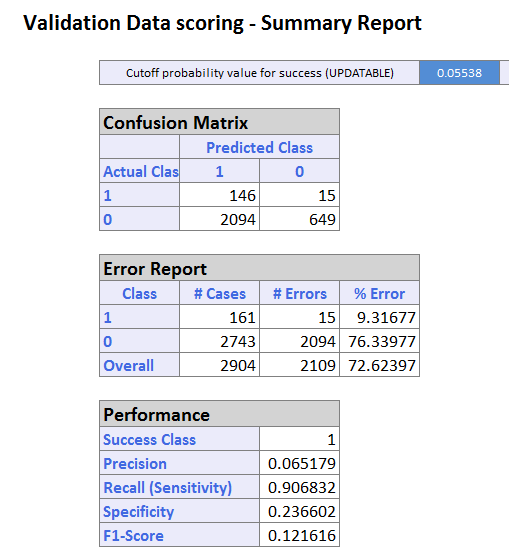
* 1. In the lift chart for testing dataset, the curve is along with straight line which indicates randomly classify the outcome by using average. The Decil-wise lift chart also do not show higher head or lower tail. Both lift chart indicate the classification model performs no better than a random classification model.



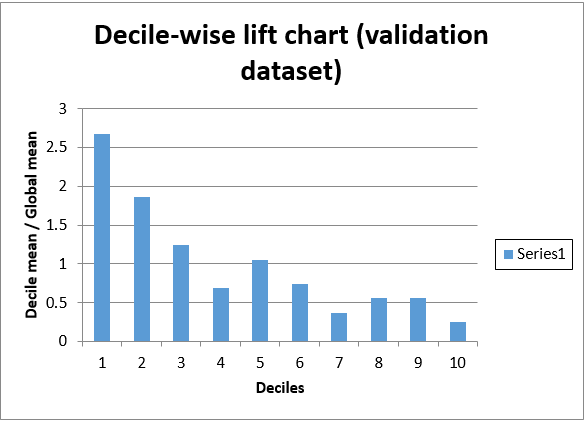
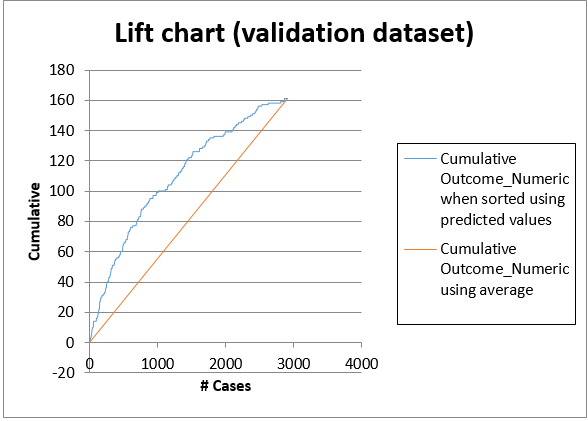
* 1. Due to the 100% error rate on one side of classified outcome, ROC Curve cannot be generated. AUC = 0.5, which indicates the model is no better than a random classifier.



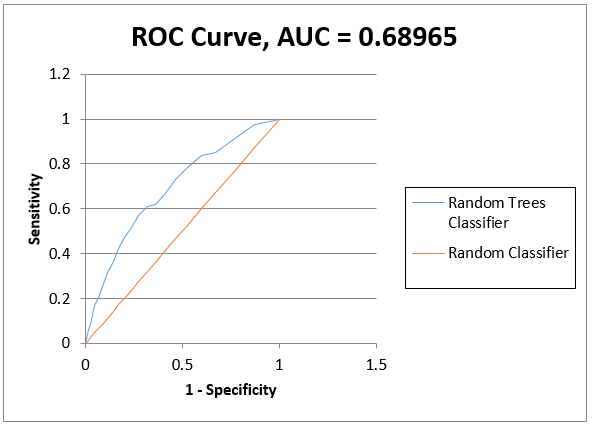
1. Random Forest
   1. In the confusion matrix of training dataset, the error rate for defaulter is 1.24% and for non-defaulter is 76%, while in the testing dataset, the error rate of defaulter increased to 9.31% and for non-defaulter remains 76%. The overall error rate for training and testing are both around 72%

* 1. In the lift chart for testing dataset, the curve is above the straight line which indicates a random classifier. The Decil-wise lift chart also has higher head and lower tail. Both lift chart indicate the classification model performs better than a random classifier.



* 1. Roc Curve for testing dataset is above the straight line which indicates random classification rules. AUC = 0.68965, which also indicate the logReg Classifier performs better than a random classifier.



1. Model Comparison:
   1. CART has the lowest overall error rate in the testing dataset. However, it failed to identify any defaulter, the performance is no better than a random classifier. Therefore, it is eliminated from the candidates of the classification models.
   2. Logistic regression has the lower overall error rate than random forest. However, the error rate of classifying a defaulter is much lower using random forest.
   3. AUC value and lift charts indicate random forest has a better performance than logistic regression.

2.4 Model selection and recommendation

1. Model selection:

In order to perform model selection, the following assumption need to be made based on the operation of bank and financial institution (lender) in real life.

Assumption 1: lender will purchase secondary insurance for any borrowers who has been classified as a potential defaulter in order to avoid their potential loss due to a mortgage default.

Assumption 2: the premium payment of purchasing a secondary insurance is much lower than the cost of missing a defaulter in classification. Assume the annual premium rate for a Lenders mortgage insurance is in the range of 0.66% to 0.75%[[2]](#footnote-1).

Assumption 3: the average mortgage payment duration is around 10 years. Defaults normally happen during after the 5th year of mortgage. The assumption is made conservatively since economic and unemployment rate changes significantly in a 10-year period, according to Federal Reserve Bank, during economic crisis, the default rate may increase to over 10%[[3]](#footnote-2).

1. Model Selection

With the three assumptions above. Random Forest has been chosen as the recommended classification model for this scenario. The following reasons are presented for reference:

* 1. Comparison has been made between Logistic Regression and Random Forest by calculating the expected costs based on the error rates of both models (testing + training).

The expected cost for Logistic Regression is

0.75%\*(102/2+871+1307+164/2)\*10+59+77-164-102 = 43.325

The expected cost for Random Forest is

0.75%\*(146/2+2094+3115+238/2)\*10+15+3-146-238 = 39.075

As a result, the expected cost of using Random Forest is lower than Logistic regression with fairly conservative assumptions[[4]](#footnote-3)

* 1. Default rate may increases due to external factors like economy downhill and unemployment rate increases.

1. Recommendation
   1. Considering the current resources and available options, we recommend Random Forest as the best Classification models to identify mortgage defaults.
   2. In order to make comprehensive decisions and finding the best model. Other models such as boosting trees should be considered, tested and compared to the current Random Forest Model.
   3. Once new data and factors are collected, both model and expected cost calculation should be refined and updated for future use.

1. https://en.wikipedia.org/wiki/Loan-to-value\_ratio [↑](#footnote-ref-0)
2. https://en.wikipedia.org/wiki/Lenders\_mortgage\_insurance [↑](#footnote-ref-1)
3. https://research.stlouisfed.org/fred2/series/DRSFRMACBS [↑](#footnote-ref-2)
4. The highest premium rate has been applied. [↑](#footnote-ref-3)