Mortgage Delinquency

Cohort 16 Capstone Project

Certificate of Data Science at Georgetown University School of Continuing Studies.

Table of Contents

[**Team members:** 2](#_Toc27073685)

[**Introduction** 2](#_Toc27073686)

[**Hypothesis and Application** 2](#_Toc27073687)

[**Domain:** Mortgage Delinquency/Default Rate Data 2](#_Toc27073688)

[**Hypothesis:** 2](#_Toc27073689)

[Business Problem 3](#_Toc27073690)

[Application: 3](#_Toc27073691)

[**Project Pipeline** 4](#_Toc27073692)

[System Design 4](#_Toc27073693)

[**Data Ingestion** 5](#_Toc27073694)

[**Wrangling** 5](#_Toc27073695)

[**Exploratory Analysis** 6](#_Toc27073696)

[**Modeling and Analysis** 8](#_Toc27073697)

[**Results** 9](#_Toc27073698)

[**Conclusion** 9](#_Toc27073699)

[**Business Application** 9](#_Toc27073700)

[**Potential future work** 9](#_Toc27073701)

[Bibliography 10](#_Toc27073702)

# **Team members:**

* Joseph Welton (Team Spokesperson)
* Jitendra Patel
* Nathaniel Rice
* Manish Pandey
* Terry Tsao

# **Introduction**

* Freddie Mac purchases loans from a servicer and charges a servicing fee (participation certificate) PC
  + Guarantees Payments to investors (Earns Guarantee Fee Fixed Interest Rate
  + Investors get a Bond (PC)
  + Through the PC, Default loans (Delinquency Mortgages are guaranteed to be paid by Freddie Mac)
* Business value of a machine learning model to determine the probability of a mortgage delinquency with a specified timestamp of the performance of the mortgage.
* This would allow a strategic business operation in data-driven decisions.
  + Allow a data driven determination on; (What loans are considered a good investment? What loans are considered a bad investment?)
* The model will allow a more tactical approach in what loans are selected and assist in driving policy changes.

# **Hypothesis and Application**

## **Domain:** Mortgage Delinquency/Default Rate Data

## **Hypothesis:**

To take key performance indicators and create a predicative model from the Freddie Mac origination data file that will allow a prediction of the default or delinquency of a mortgage within a 60-80% accuracy. The sample of mortgages used were from the Freddie Mac origination data file were the first payment of the mortgage occurred after December 2011.

**Project Description:**

* Mortgage Delinquency/Default rate using key performance indicators (KPI) to build a data frame model which will be used for predictive analytics to determine delinquency/default of the mortgage along with current state of the economy.
* Utilizing statistical trend and regression analysis and methodologies to test the model. Utilizing test data to test the model to accurately forecast the probability that a mortgage with become delinquent/default.
* Present the results of the model indicating the accuracy of the model and the visualization of results.

## Business Problem

To be able to develop a predictive model using origination data received from the service of the loan to determine the performance of the loan at the follow periods.

* Within the first year of the loan
* Within the first three years of the loan
* Within the first five years of the loan

The dataset utilized for the Mortgage Delinquency model does not have any demographic data which could be utilized for bias determination. However, with adopting new technology to allow a new decision-making process a cautious approach may be needed.

Though some of the data features such as credit score and income to debt ratio is within the dataset which could be utilized to make bias determinations. Other features may also be used for biased determinations though the exploratory analysis may not or may not support bias actions.

Therefore, the previous recommendation of a cautious approach would be the best action until the decision make process is stabilized and the model can be expanded to more complex decision determinations.

## Application:

The predictive model will present leadership with information that will drive a decision. This data driven decision will act as an internal control option for determining policy determinations regarding loan sellers and the participation fee. The model will also allow an adaptive business process in determinizing the probability of a loan becoming delinquent and the specified intervals.

# **Project Pipeline**

## System Design

# **Data Ingestion**

Extracted the data from the Freddie Mac website and concatenated the data into two files. The origination file which contained all the data features of the loan at loan inception. The performance file contains the features deemed valid to measure the performance of the loan on monthly basis. These data files where then ingested into a PostgreSQL database from which SQL code is written to extract applicable data queries for any business and exploratory need.

We downloaded the public datasets from Freddie Mac. We concatenated the multiple year data sets into monthly performance data and Origination Data. The Monthly Performance data contained the data regarding the performance of the loan by month from the inception of the loan. The Origination Table has all the initial key data elements for the loan. The Origination Table and the Monthly Performance table are linked by a Loan Sequence Number. We have over 8 million loans with their performance data to wrangle. There are 27 features in the origination table, 26 features in the performance table. Therefore, we have over 53 features per loan.

A Postgres Database was setup with Origination Table and Monthly Performance Table. Each member of the team hosts their own local version of the database for exploratory data analysis. The initial database set up is for the team to conduct exploratory data analysis. However, once we determine and settle on the dataset, a designated team member will be the main database controller so we can write once and ready many (WORM) with the data.

We extracted the final dataset by taking a sample of the larger dataset. We choose a random sample utilizing a random state sampling 18% of the main dataset. This yielded a sample of 1,474,944 loans with all the applicable features needed for the analysis. We deemed this a valid sample size based on statistical sample size calculation with a confidence level of 95% and confidence interval (margin of error) at .08.

# **Wrangling**

We reviewed the data under the Four Vs of Big Data: Volume (Scale of the Data), Variety (Diversity of Data), Veracity (Certainty of Data), and the Velocity (Speed of Data). The combined dataset from Freddie Mac has over 8 million loans in the Origination Table and over 10 years of Performance data for each loan. We combined/joined tables from the Origination Table and the Performance Table by the Loan Sequence number. This allowed us to obtain a Variety of Loan Data with various aspects of Origination data and performance data. We were able to use SQL code to extract a variety samples from the Postgres Database. This variety of data will allow us thru exploratory data analysis and feature analysis allowed us to determine which key performance indicators/features should be utilized for modeling purposes.

The final dataset is framed with the data features from the Origination Table with the Delinquency status from the performance table. We extracted the data file as a pickle file for easy use among the team. From the final pickle file where can breakdown the dataset for exploratory data analysis and applicable machine learning actions via the data science pipeline.

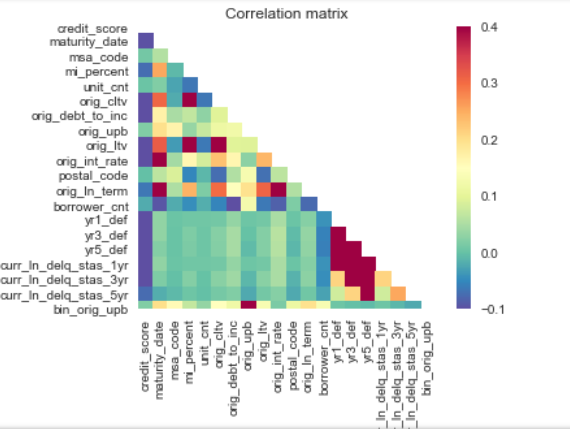
The performance of the loan is broken down into three sections:

1. The first section is if the loan is delinquent within the first year (0-12 months) from the origination date of the loan based on the loan delinquency status from the performance data table.
2. The second section is if the loan is delinquent within the first three years (0-36 months) from the origination date of the loan based on the loan delinquency status from the performance data table.
3. The third section is if the loan is delinquent within the first five years (0-72 months) from the origination date of the loan based on the loan delinquency status from the performance data table.

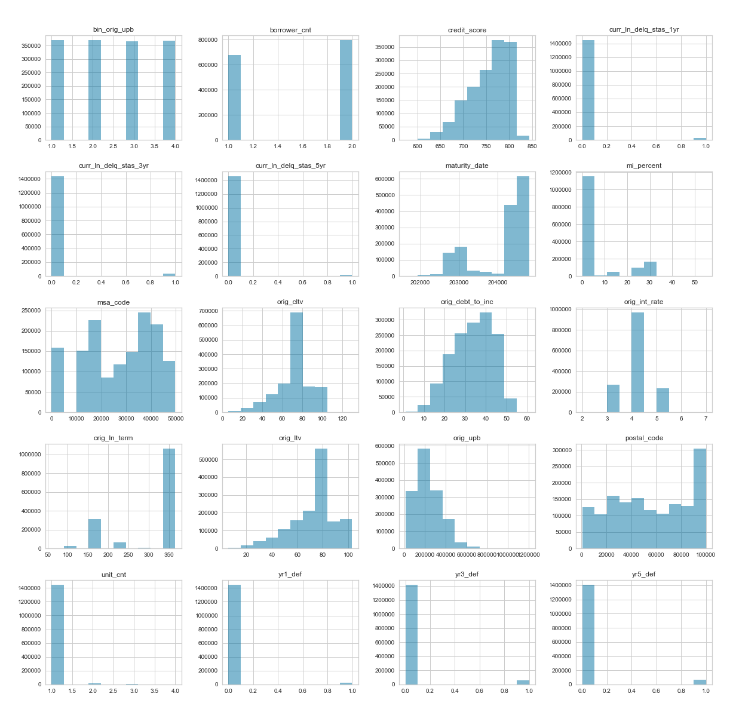
The machine learning target is the second section/instance. Utilizing the data within the origination data table to determine the loan delinquency status within the first three years or within the first 36 months of the loan.

# **Exploratory Analysis**

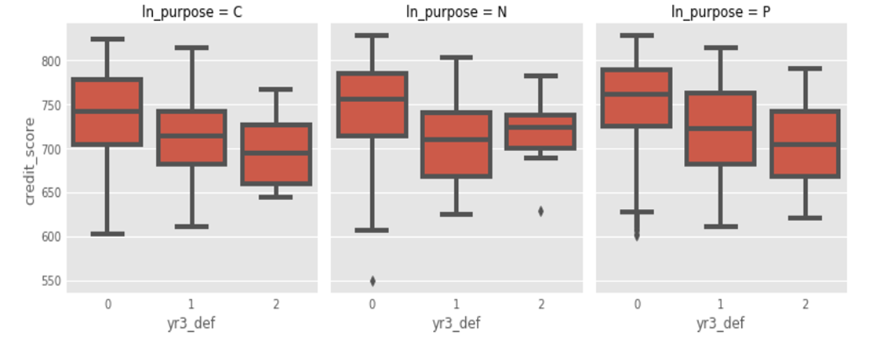
We used Jupyter Notebook for the principal python platform for our exploratory data analysis. We used the read CSV Feature from Pandas for majority of our exploratory data analysis. Once we had the data in a Pandas data frame in python, we continued the exploratory analysis. We found that majority of the data features in the Origination Table were categorical data. Some of which could be transferred into a Boolean value. Utilizing Scikit-learn, NumPy, Matplotlib, Seaborn, and Pandas for feature analysis and statistical review.

Correlation Matrix Graph shows which features are correlated with one another. 

* Histograms of the numeric features within the dataset: Utilized for Exploratory data analysis for understanding the data features within the dataset.



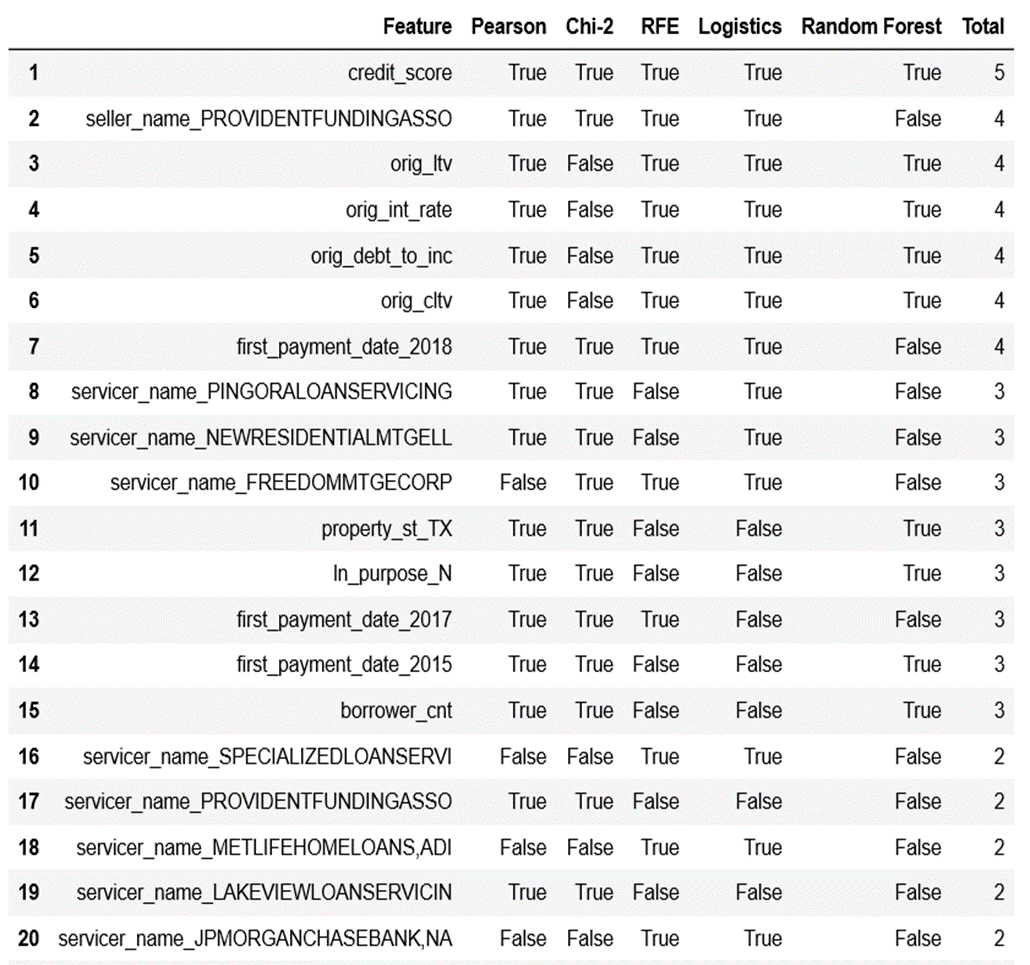
* Boxplot showing the loan purpose via delinquency status within year 3 from the mortgage inception:
  + [ C = Cash-Out Refinance, P = Purchase, N = No Cash-Out Refinance]

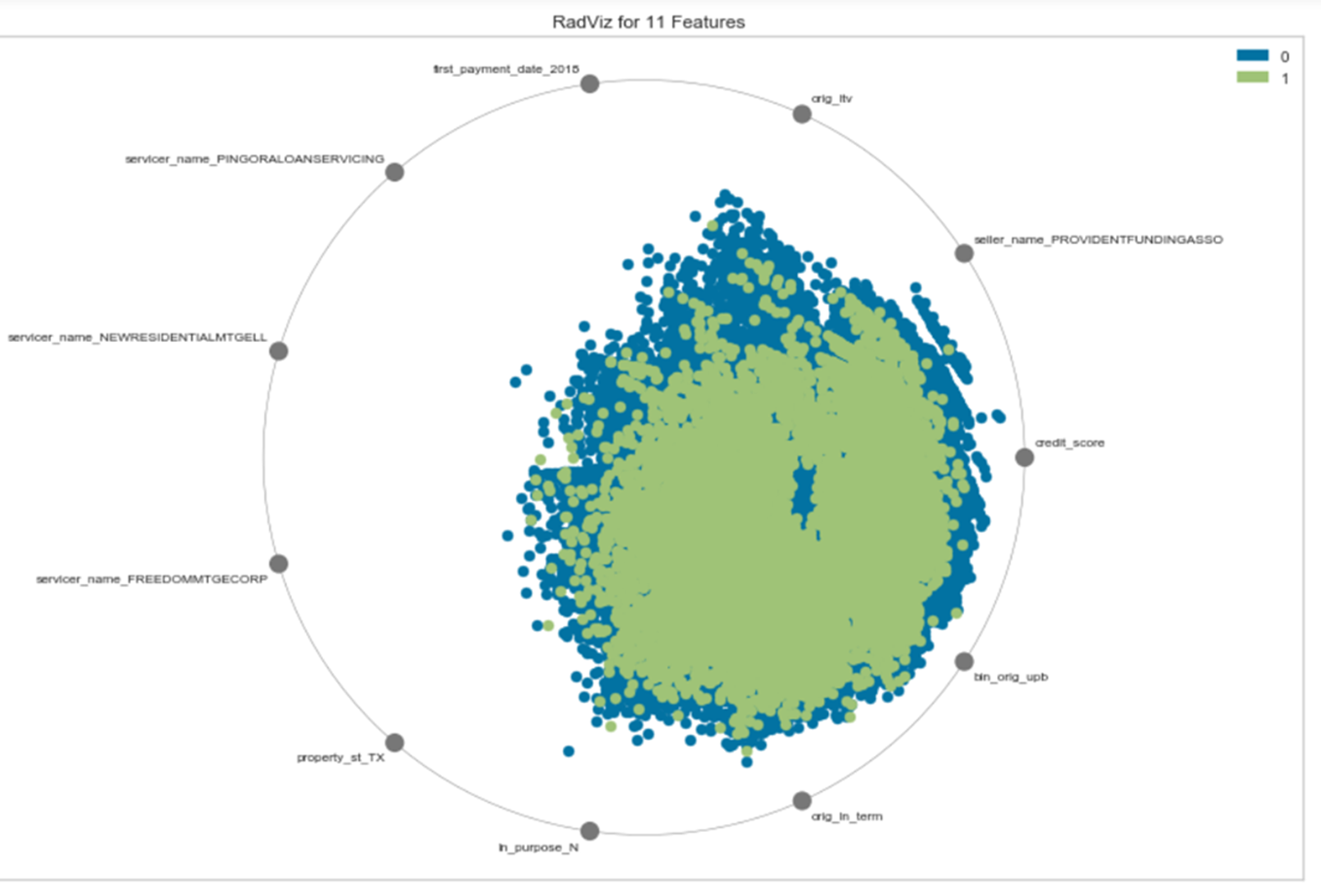


# **Modeling and Analysis**

Based on the correlation matrix from the exploratory data analysis and the feature analysis we determined which features appear to affect the target of mortgage delinquency within 3 years. The target is an instance of mortgage delinquency at any point during the first 3 years or first 36 months of the mortgage.

Feature Analysis:



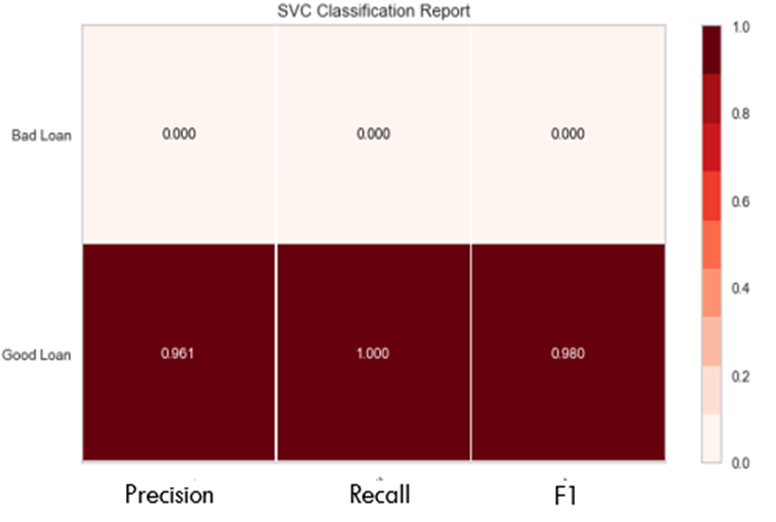


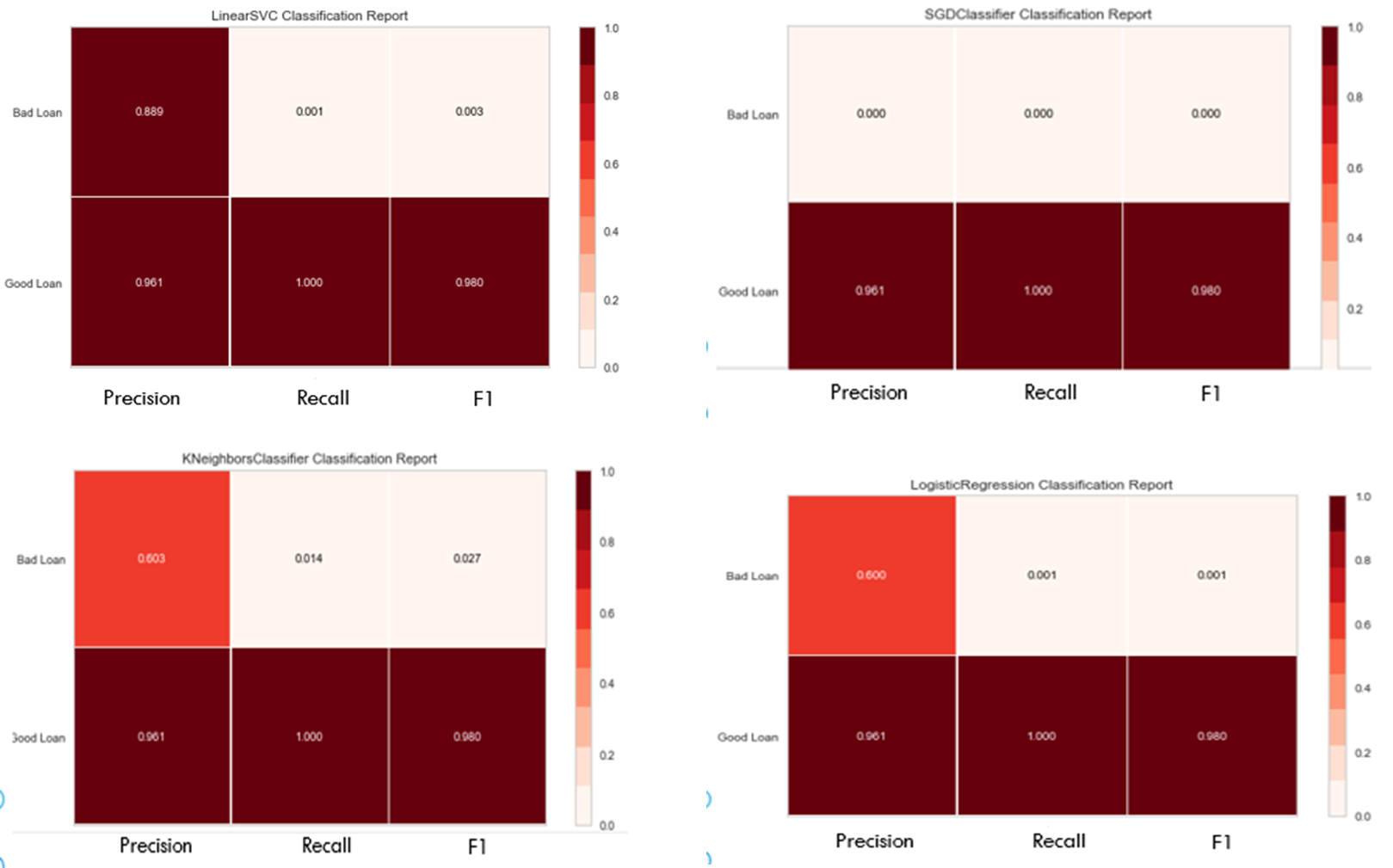
From the feature analysis we determined which data features are ingested into the model for analysis.

We executed the following models:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Loan Classification** | **Precision** | **Recall** | **FI Score** |
| SVC | Bad Loan | 0.000 | 0.000 | 0.000 |
| SVC | Good Loan | 0.961 | 1.000 | 0.980 |
| Linear SVC | Bad Loan | 0.889 | 0.001 | 0.003 |
| Linear SVC | Good Loan | 0.961 | 1.000 | 0.980 |
| SGD Classifier | Bad Loan | 0.000 | 0.000 | 0.000 |
| SGD Classifier | Good Loan | 0.961 | 1.000 | 0.980 |
| K Neighbors Classifier | Bad Loan | 0.603 | 0.014 | 0.027 |
| K Neighbors Classifier | Good Loan | 0.961 | 1.000 | 0.980 |
| Logistic Regression | Bad Loan | 0.600 | 0.001 | 0.001 |
| Logistic Regression | Good Loan | 0.961 | 1.000 | 0.980 |
| Logistic Regression CV | Bad Loan | 0.000 | 0.000 | 0.000 |
| Logistic Regression CV | Good Loan | 0.961 | 1.000 | 0.980 |
| **Bagging Classifier** | **Bad Loan** | **1.000** | **0.735** | **0.847** |
| **Bagging Classifier** | **Good Loan** | **0.980** | **1.000** | **0.995** |
| Extra Trees Classifier | Bad Loan | 1.000 | 1.000 | 1.000 |
| Extra Trees Classifier | Good Loan | 1.000 | 1.000 | 1.000 |
| Random Forest Classifier | Bad Loan | 1.000 | 0.980 | 0.990 |
| Random Forest Classifier | Good Loan | 1.000 | 1.000 | 1.000 |

* Heat Map visualizations of the Classification Models that show the Precision, Recall and F1 score





# **Results**

The Bagging Classifier Model was the only model that supported the hypotheses. We hypothesized that we could created a model with 60%-80% accuracy score. The Bagging Classifier Model was 84.7% accurate which supports the hypotheses. We also had a Recall score of 73.5% for the Bagging Classifier Model. This indicates that the accuracy to properly identify a delinquent loan when a delinquent loan was selected from the dataset.

# **Conclusion**

The results of the Bagging Classifier Model are promising however, since the other models were overfitting and/or underfitting indicate there might be other issues within the dataset.

The main aspect would be that Freddie Mac has already vetted the data and deem which loans were a good investment. The key performance indicators derived from the initial vetting process are the data features in the Origination Data Set and in the Performance Data Set.

More sample datasets would need to run to order to come to a more complete conclusion that the model has a distinct business value that will support data driven determinations.

# **Business Application**

The results of the model support expansion of the models for further exploration for data driven decision. This will other models to be applied along with change any data features to generate a more consistent and accurate model.

# **Potential future work**

The performance data of the mortgages is captured on a monthly basis. This can be reviewed as a multivariate time-series. For future exploration time-series models should be explored to determine trends within the data features of the monthly performance data. Some of the models/testing options to investigate would be Dickey-Fuller, and ARIMA (Autoregressive Integrated Moving Average).

Other option would be to change the target from delinquency status to the income to debt ratio from the origination dataset. This change could show different indicators related to the what affect the debt to income has on the probability of a mortgage having an instance of delinquency within the performance of the loan.

# Bibliography

Agency, F. H. (2016). *TREND: Federal Housing Finance Agency. House Price Index: House Price Index - All Transactions, 1975/1 - 2016/1. Data-Planet™ Statistical Datasets by Conquest Systems, Inc. Dataset-ID: 057-001-001*. Retrieved 10 18, 2019, from https://search.labs.datacite.org/works/10.6068/dp157f983aa7e3

Agency, F. H. (2017). *TREND: Federal Housing Finance Agency. House Price Index: House Price Index - All Transactions | State: New York | Seasonally Adjusted: Non-Seasonally Adj, 1975/1 - 2017/2. Data-Planet™ Statistical Datasets by Conquest Systems, Inc. Dataset-ID: 057-001-001*. Retrieved 10 18, 2019, from https://search.datacite.org/works/10.6068/dp15f08b7a62e41

Askitas, N., & Zimmermann, K. F. (2011). *Detecting Mortgage Delinquencies*. Retrieved 10 18, 2019, from http://ftp.iza.org/dp5895.pdf

Davis, M. A., Larson, W. D., Oliner, S. D., & Smith, B. (2019). *Mortgage Risk Since 1990*. Retrieved 10 18, 2019, from https://econpapers.repec.org/repec:hfa:wpaper:19-02

*HUDUser.gov - HUD USER*. (n.d.). Retrieved 10 22, 2019, from http://www.huduser.org/portal/

Liu, H. (2017). *Essays in Risk Management and Financial Econometrics*. Retrieved 10 18, 2019, from https://escholarship.org/uc/item/77r9f9z6

Ma, C. (2016). *House Price Expectations and Mortgage Default Decisions*. Retrieved 10 18, 2019, from https://papers.ssrn.com/sol3/delivery.cfm/ssrn\_id2831047\_code2494459.pdf?abstractid=2709563&mirid=1

Patrabansh, S. (2013). *A Study of First-Time Homebuyers*. Retrieved 10 18, 2019, from https://papers.ssrn.com/sol3/papers.cfm?abstract\_id=2533963

<https://www.investopedia.com/terms/r/realestateowned.asp>

<https://www.fhfa.gov/DataTools/Downloads/Pages/Public-Use-Databases.aspx>

<https://www.federalreserve.gov/data/mdrm.htm>

<https://www.federalreserve.gov/data.htm>

<https://www.fanniemae.com/portal/funding-the-market/data/loan-performance-data.html>

<https://fred.stlouisfed.org/categories/32440>

<https://www.consumerfinance.gov/data-research/mortgage-performance-trends/download-the-data/>

<https://www.consumerfinance.gov/data-research/hmda/historic-data/>