**Mortgage Delinquency Team Progress Report 2: November 16, 2019**

**Team:** 

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

**Hypothesis:** To take key performance indicators and create a predicative model that will allow an accurate prediction of the default or delinquency rate of a mortgage within a 60-80% accuracy.

**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.

**Meetings**: The team had meetings on Wednesdays during the week to discuss the current project status thru Go-To-Meeting. Other ad hoc meetings were conducted at the Reston Regional Library for in person meeting discussions. The team also would have lunch meetings during Saturday classes along with meeting after class when needed. Pending individual availability enough quorum attended the applicable meetings. Progress actions were communicated via email and/or phone calls for those members that could not attend meetings.

**Classes: With Actions and Tasks**

* Statistics
  + Progress Report I
  + Pandas and NumPy
    - Data Exploration – Correlation and descriptive statistics
* Machine Learning
  + Wrangling the data for various machine learning techniques to determine which is the best model for the data set and business problem.
* Visual Analytics
  + Progress Report II
* Applied Data Analytics
  + Demo and Final Report

**Task Actions:**

|  |  |  |
| --- | --- | --- |
| **Task** | **Task Done by** | **Notes** |
| Capstone Proposal | Joseph Welton |  |
| Architecture | Manish Pandey |  |
| Progress Report I | Joseph Welton |  |
| Database Research | Joseph Welton, Jitendra Patel, Nathaniel Rice, Manish Pandey, Terry Tsao |  |
| PostgreSQL Database Setup | Jitendra Patel, Nathaniel Rice, Terry Tsao | PostgreSQL/Python |
| PostgreSQL Database Data Dump | Jitendra Patel, Nathaniel Rice, Terry Tsao | PostgreSQL/Python |
| PostgreSQL - Data Extract: SQL Code actions | Joseph Welton, Jitendra Patel, Nathaniel Rice, Manish Pandey, Terry Tsao |  |
| Data Research-Locating Data Source(s) | Joseph Welton, Jitendra Patel, Nathaniel Rice, Manish Pandey, Terry Tsao | Freddie Mac Data |
| Data Exploration-Pandas | Joseph Welton, Jitendra Patel, Nathaniel Rice, Manish Pandey, Terry Tsao | Freddie Mac Data |
| Data Graphic Exploration -Python Pandas & Matplotlib | Joseph Welton, Jitendra Patel, Nathaniel Rice, Manish Pandey, Terry Tsao | Freddie Mac Data |
| Data Statistics Exploration -Python Pandas | Joseph Welton, Jitendra Patel, Nathaniel Rice, Manish Pandey, Terry Tsao | Freddie Mac Data |
| Algorithm Research | Manish Pandey | Scikit Learn Exploration |
| Modeling and Application | Joseph Welton, Jitendra Patel, Nathaniel Rice, Manish Pandey, Terry Tsao |  |
| Pushing actions to GIT Hub | Joseph Welton, Jitendra Patel, Nathaniel Rice, Manish Pandey, Terry Tsao |  |
| Scikit-Learn | Joseph Welton, Jitendra Patel, Nathaniel Rice, Manish Pandey, Terry Tsao |  |
| Reporting and Visualization | Joseph Welton, Jitendra Patel, Nathaniel Rice, Manish Pandey, Terry Tsao |  |
| Yellow Brick | In progress |  |
| Flask | In progress |  |
| Spark | In progress |  |
| Progress Report II | In progress |  |
| Demo and Final Report | In progress |  |

**Questions and actions discovered during analyzing Data:**

Purpose:

Freddie Mac purchases loans from Sellers and charges the Sellers a guarantee fee on the loans. This is the primary method where Freddie Mac makes money. The ability for Freddie Mac leadership to make more informed determinations regarding the loans that are purchased will drive development of policy and internal controls.

* 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.

# **Hypothesis and Application**

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

## **Hypothesis:**

To take key performance indicators and create a predicative model that will allow an accurate prediction of the default or delinquency rate of a mortgage within a 60-80% accuracy.

(The data elements/indicators on the Origination Table are deemed to be the most important factors when determining the monthly performance of the loan.)

## Application:

The model will allow improved investment decisions permitting leadership to choose which loans are a better option for a Freddie Mac purchase and determine if the loans already purchased were a reliable investment. The loans reviewed are less than one year old, these loans are ingested into the Linear Regression Machine Learning Model to indicate a likelihood of becoming delinquent within the first year. This likelihood allows leadership to make a more informed determination regarding purchase options for investment purposes. However, there are risks that drive where a loan could potentially become delinquent within year 3, year 5, or after 10 years. These loans ingested into the model would also indicate leadership decisions that resulted in the purchase of a poor performing loan. Analysis of year 3, year 5 and after 10 years would need to be explored further to determine what seller/loan servicer sold the loan to Freddie Mac then further discussions regarding changes to internal controls and policy actions to protect the investment integrity of Freddie Mac.

# **Data Ingestion**

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.

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.

# **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 had 5 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 determine which key performance indicators/features should be utilized for modeling purposes.

Initial exacts from the database were downloaded into a Comma Separated Values (CSV) files. These CSV files were then loaded into a python data frame for exploratory analysis.

# **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.

# Bibliography

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

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

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

Initial Data Sources:

<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/>