Pundir, Saurabh

saurabhpundir.data@gmail.com

Mortgage CUSTOMER SEGMENTATION Project

Capstone Project Summary

Table of Contents

[**Problem** 1](#_Toc516405276)

[**Who might care?** 2](#_Toc516405277)

[**Data** 2](#_Toc516405278)

[**Approach** 3](#_Toc516405279)

[I. Step I- Data Extraction: 3](#_Toc516405280)

[II. Step II – Data Wrangling: 3](#_Toc516405281)

[III. Step III – Exploratory data analysis: 3](#_Toc516405282)

[IV. Step IV – Machine Learning: 3](#_Toc516405283)

[**Project Deliverables** 4](#_Toc516405284)

# **Problem**

The mortgage industry has done $1.4 trillion of new mortgages in 2016. The top three lenders have a total share of $150 million, with the leader having a share of $80 million. Rest of the pie is distributed amongst 400 small and medium-sized lenders.

According to the problem defined by top players, the mortgage companies have not been able to understand their target customers and provide suitable loan products.

As per a recent study conducted by J.D. Powers, 63% of customers would leave their mortgage servicer for better customer service. The same study shows that 27% of first-time buyers and 21% of all borrowers regret their choice of lender.

Mortgage companies want to increase their market share and for doing so they need to understand their customers better. This project aims to use data from Consumer Financial Protection Bureau and build an unsupervised machine learning model to segment their customer base.

# **Who might care?**

Mortgage companies like Quicken Loans, Wells Fargo, Chase and credit unions, etc. can use such a model to understand their customer base. They can create products catering to a segment, decide marketing allocation, and formulate the strategy for reaching out the segmented customers in a unique way.

The data also has features like the type of homes, size of family and features related to clients etc. This segmentation can also provide insights for Real Estate companies as every Mortgage customer is also a Real Estate customer.

# **Data**

The data will be acquired from the Consumer Financial Protection Bureau. In this project, we will mainly focus on the data collected during 2015-2017. There is a possibility to use census data to gather more demographic information.

The data is available in CSV and similar delimiter format. During our study, we will consider the below-mentioned data fields.

|  |  |  |
| --- | --- | --- |
| Data Field Type | Valid Values | Descriptions and Examples |
| 1 | Record Identifier - Value is 2 | Numeric |
| 2 | Respondent-ID | Alphanumeric |
| 3 | Agency Code | Numeric |
| 4 | Loan Type | Numeric |
| 5 | Property Type | Numeric |
| 6 | Loan Purpose | Numeric |
| 7 | Owner Occupancy | Numeric |
| 8 | Loan Amount | Numeric |
| 9 | Preapprovals | Numeric |
| 10 | Type of Action Taken | Numeric |
| 11 | Metropolitan Statistical Area/Metropolitan Division | Alphanumeric |
| 12 | State Code | Alphanumeric |
| 13 | County Code | Alphanumeric |
| 14 | Census Tract | Alphanumeric |
| 15 | Applicant Ethnicity | Numeric |
| 16 | Co-applicant Ethnicity | Numeric |
| 17 | Applicant Race: 1 | Numeric |
| 18 | Applicant Race: 2 | Numeric |
| 19 | Applicant Race: 3 | Numeric |
| 20 | Applicant Race: 4 | Numeric |
| 21 | Applicant Race: 5 | Numeric |
| 22 | Co-applicant Race: 1 | Numeric |
| 23 | Co-applicant Race: 2 | Numeric |
| 24 | Co-applicant Race: 3 | Numeric |
| 25 | Co-applicant Race: 4 | Numeric |
| 26 | Co-applicant Race: 5 | Numeric |
| 27 | Applicant Sex | Numeric |
| 28 | Co-applicant Sex | Numeric |
| 29 | Applicant Income | Alphanumeric |
| 30 | Type of Purchaser | Numeric |
| 31 | Denial Reason: 1 | Numeric |
| 32 | Denial Reason: 2 | Numeric |
| 33 | Denial Reason: 3 | Numeric |
| 34 | Rate Spread | Alphanumeric |
| 35 | HOEPA Status | Numeric |
| 36 | Lien Status | Numeric |

# **Approach**

## Step I- Data Extraction:

A large dataset containing loan information is available as a yearly CSV file. We will extract data from 2015-2017. The idea is to consolidate the data in one singular file.

## Step II – Data Wrangling:

All the missing fields will be identified and will be handled. If required data will be cleaned for null values or empty fields. The data conversion will be performed if required. Furthermore, columns will be extracted for exploratory data analysis and machine learning. Data will be stored in a pickle file for further processing.

## Step III – Exploratory data analysis:

Various features in data will be explored through graphs and charts to get more insights into the data. The inferential statistics will be performed to understand data.

## Step IV – Machine Learning:

Unsupervised Machine learning algorithms will be applied for clustering and segmentation.

# **Project Deliverables**

The project deliverable will contain the following:

1. This document explains the approach of the project.
2. Juypter notebooks containing python code for steps mentioned in approach section i.e. Data extraction, wrangling, exploratory analysis, machine learning.
3. Final report of the project.
4. Presentation slide deck created on this project.