



CMPE-255 Data Mining

Final Project Report

HDMA Washington State Home Loans, 2016

Fall - 2021

INSTRUCTOR - Prof. Jorjeta Jetcheva

Team

Ishan Hitesh Malkan: 015281968

Romil Viral Shah: 015250599

Pallav Parikh: 015275481

1. Introduction

Analysing home loan related dataset of the Washington State from 2016. It contains various columns such as what was the final outcome of the loan application and various details about the application such as income, gender, race of the applicant and area related variables. The dataset was taken from Kaggle.

1.1 Motivation

There are several motivations behind this project. As almost every human being aims to buy a house at least once in his/her lifetime, it is crucial to understand the dynamics behind what can be the deciding factors for your home loan application to be accepted or rejected as well as the amount of interest rate you are able to get. Apart from this, there are various talks going around regarding the equality based on gender, race etc. and it would be interesting to analyse whether any significant effects these types of variables have on the loan application decision, if yes then what might be the reasons.

1.2 Objective

In this project there are 2 main objectives, the first one is to analyse and extract as much information as possible about the various factors affecting the loan application decision and analyse the trends and patterns found in the data among its variables. The second objective is to determine whether the dataset is good enough to make a predictive model to classify whether the application would be approved or not.

1.3 Approach

The problem is approached by first cleaning the dataset by removing the useless fields and processing the remaining fields such that meaningful information can be extracted from the same. Once the dataset is well processed then look at the trends of the dataset and relationship among its columns and rows. Finally after selecting the most useful features which are most related to the loan application decision, a modelling exercise is tried to see whether the columns are able to predict the loan application decision.

1.4 Literature review

As it is advised to perform detailed literature review before starting any new project, this project was no exception. In any Data Science project there are two main aspects of literature review that are the domain knowledge and the relevant data science fields knowledge. An extensive literature review was made on both the aspects. For the domain knowledge, various terminologies were explored such as HMDA (Home Mortgage Disclosure Act) etc. detailed information about which can be gathered from

the links in the references. Secondly for the technical part, various data pre processing and cleaning methods were explored along with classification models and python programming tools. The content covered in the class was mostly enough for the technical literature review as the concepts learnt in the class were mostly used in implementation of this project.

2. System Design and Implementation

This section talks about the various algorithms considered, tools and technologies used, and other implementation details.

2.1 Algorithms considered

Many different algorithms were considered for both the data processing and the modelling part. Starting from data processing, for missing value imputation the algorithms considered were imputing them with median, mode or creating a separate category for them. Then for binning the data a decision tree regressor algorithm was used along with cross tables and histograms for visualization of the data. Some other algorithms used to process the data were one hot encoder, label encoder etc. to make the dataset in proper format to input in the models.

For the modelling part, various classification algorithms were explored and implemented such as logistic regression, gradient boosting classification etc. along with hyper parameter tuning techniques. As we will see in the later sections the gradient boosting classification algorithm works better compared to others for this dataset.

2.2 Technologies and tools used

Python programming language was used for the implementation of this entire project. Some important libraries that were used are: pandas and numpy libraries to perform data transformation and data processing on tabular data (matrices), matplotlib and seaborn libraries to visualize the data into graphical format, and finally the scikit-learn library for all the other tasks of data transformation, modelling, evaluation metrics etc.

Anaconda tool was used to create the python environment and the code was written in Jupyter Notebook. Excel was used for basic tabular data visualization and saving the data in comma separated formats.

2.3 Data Flow

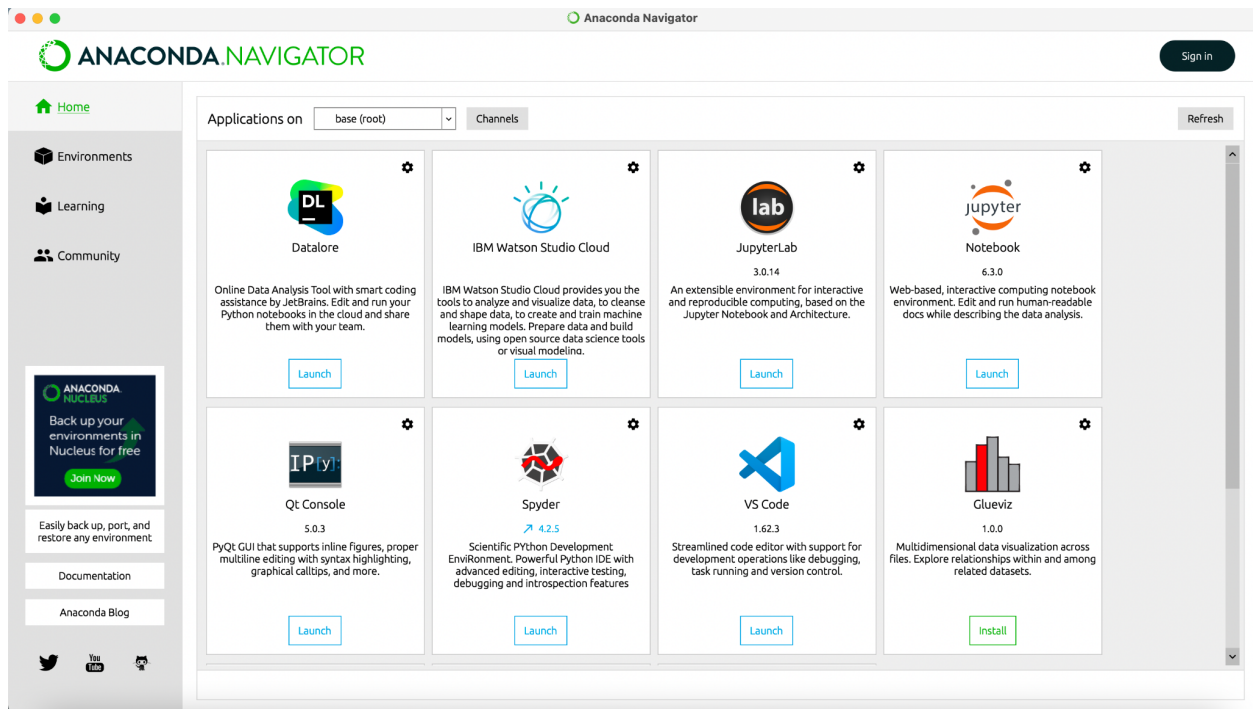
As the dataset was received directly from kaggle in comma separated variables format, with all the required columns, work was started directly using that data and no other external data source was used. This single dataset was flowed into the python dataframe using pandas library and all the processing was done, using the tools mentioned in the above section.

2.4 Screenshots

Below are some of the screenshots from the project. The screenshots are of the tools used as there is no GUI in this project.

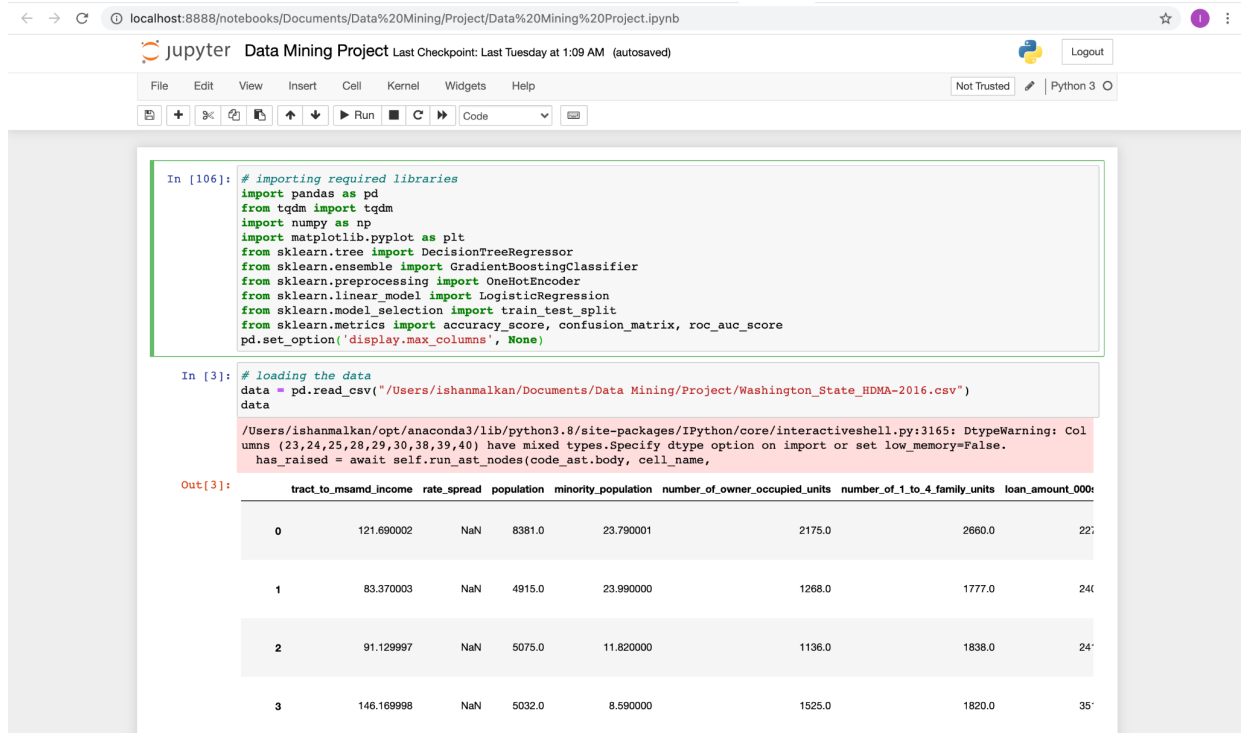
Anaconda Navigator:

Anaconda comes with almost all the required dependencies of data science projects and the main page of anaconda looks like the below screenshot.



Jupyter Notebook:

A screenshot of the code from the project in Jupyter Notebook



```
In [106]: # importing required libraries
import pandas as pd
from tqdm import tqdm
import numpy as np
import matplotlib.pyplot as plt
from sklearn.tree import DecisionTreeRegressor
from sklearn.ensemble import GradientBoostingClassifier
from sklearn.preprocessing import OneHotEncoder
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score, confusion_matrix, roc_auc_score
pd.set_option('display.max_columns', None)
```

```
In [3]: # loading the data
data = pd.read_csv("/Users/ishanmalkan/Documents/Data Mining/Project/Washington_State_HDMA-2016.csv")
data
```

Users/ishanmalkan/opt/anaconda3/lib/python3.8/site-packages/IPython/core/interactiveshell.py:3165: DtypeWarning: Columns (23,24,25,28,29,30,38,39,40) have mixed types.Specify dtype option on import or set low_memory=False.

```
Out[3]:
```

	tract_to_msamd_income	rate_spread	population	minority_population	number_of_owner_occupied_units	number_of_1_to_4_family_units	loan_amount_000s
0	121.690002	NaN	8381.0	23.790001	2175.0	2660.0	22
1	83.370003	NaN	4915.0	23.990000	1268.0	1777.0	24
2	91.129997	NaN	5075.0	11.820000	1136.0	1838.0	24
3	146.169998	NaN	5032.0	8.590000	1525.0	1820.0	35

3. Proof of concept evaluation

This section will talk about the datasets used, methodologies followed, visualizations using graphs and evaluation of the results.

3.1 Dataset used

The dataset was taken from kaggle and it contained 466,566 rows (unique loan applications) and 47 columns (features of each application). Following is the link to the kaggle dataset:

<https://www.kaggle.com/miker400/washington-state-home-mortgage-hdma2016>

The most interesting column of the dataset (also the target variable for this project) was the “action_taken_name” meaning what action was taken on that particular application. It had values such as “loan originated”, “Application denied by financial institution” etc. Some of the other columns were reasons for denial of loan, gender, race of applicant and co applicant, income, median income etc.

Several preprocessing tasks were performed on the dataset, starting from removing the constant columns and name columns. Then removing the columns that had > 95% null

rows because less than 5% data points are statistically insignificant. Then the categorical and numerical features were separated as different preprocessing was required on both of them.

For numerical features first the null values were replaced by median if that particular feature contained <5% nulls but if a feature contains more than 5% nulls it is not replaced by anything because it might add too much noise to the data and hence in such a case null is considered a separate category and the numerical feature is treated as categorical after binning the remaining values. The features where the nulls were replaced by median are also binned using a decision tree regressor. (put the feature as the dependent and independent variable in the tree and then prune the tree as per the requirement of the bins).

For the categorical columns, the ones having <5% nulls have nulls replaced with mode (but there are no such columns) and the ones having >5% nulls have nulls created as a separate category.

Some statistics from the dataset:

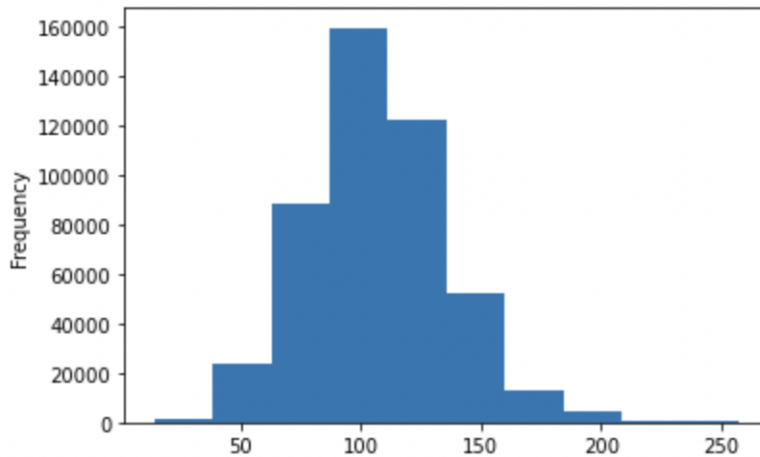
- a. Distribution of action taken name feature.

Here while modelling only the top 2 categories would be considered to create a binary classification model, if that works well then the others can be considered.

Loan originated	263712
Application denied by financial institution	64177
Application withdrawn by applicant	60358
Loan purchased by the institution	48356
File closed for incompleteness	18176
Application approved but not accepted	11735
Preapproval request denied by financial institution	35
Preapproval request approved but not accepted	17

Name: action_taken_name, dtype: int64

- b. Histogram of the feature "tract_to_msamd_income"



Several other stats and visualizations were analysed on the dataset but considering the size of this document are not added here.

3.2 Methodology followed

After the preprocessing of the data as described above, their cross tables were created with the target variable to check whether they are rank ordering or not. The features which were rank ordering well as it is (after binning using decision tree regressor or is a categorical feature) were taken as it is, the features which were not rank ordering as it is but looked like merging/splitting of some bins may create proper rank ordering were taken after rebinning them manually, and the feature which looked like they are not rank ordering at all were not considered during modelling. An example from each of the three categories are mentioned below.

- a. Feature rank ordering as it is (after binning using decision tree regressor)

Tract to msamd income

Here it is visible that as the value of feature increases the % of loan originated increases and hence this feature is rank ordering well and can be considered for the model.

action_taken_name	Application denied by financial institution	Loan originated	% Loan originated
tract_to_msamd_income			
(-2.0, 81.99]	14058	46114	0.766370
(81.99, 95.985]	12375	47626	0.793753
(95.985, 109.98]	12977	51371	0.798331
(109.98, 124.625]	11209	50927	0.819605
(124.625, 142.02]	7503	35673	0.826223
(142.02, 170.68]	4825	24774	0.836988
(170.68, 100002.0]	1230	7227	0.854558

b. Feature not rank ordering well now but will after some manual binning

Here it is visible that the feature “loan_amount_000s” does not rank order as it is but if we merge the last few bins it will rank order properly. The first image shows the bins after decision tree regressor and the second image shows bins after manually merging some of the bins

action_taken_name	Application denied by financial institution	Loan originated	% Loan originated
loan_amount_000s			
(-2.0, 197.5]	28471	79313	0.735851
(197.5, 351.5]	24239	114315	0.825057
(351.5, 633.5]	9288	58847	0.863682
(633.5, 1345.5]	1879	9511	0.835031
(1345.5, 100002.0]	300	1726	0.851925

action_taken_name	Application denied by financial institution	Loan originated	% Loan originated
loan_amount_000s			
(-2.0, 197.5]	28471	79313	0.735851
(197.5, 351.5]	24239	114315	0.825057
(351.5, 100002.0]	11467	70084	0.859389

c. Feature not rank ordering at all

Here it is visible that the population feature is not rank ordering at all

action_taken_name	Application denied by financial institution	Loan originated	% Loan originated
population			
(-2.0, 3049.0]	5325	19735	0.787510
(3049.0, 3989.5]	9844	38918	0.798121
(3989.5, 5502.5]	22219	93706	0.808333
(5502.5, 6375.0]	10868	47023	0.812268
(6375.0, 7338.5]	8591	35524	0.805259
(7338.5, 8893.0]	5542	21168	0.792512
(8893.0, 100002.0]	1788	7638	0.810312

3.3 Graphs showing different parameters/algorithms evaluated in a comparative manner, along with some supportive text

After following the above methodologies, 10 features were finalized to create models. But as shown in the examples above, even though the features rank order, the difference does not seem enough to be able to make a predictive model, the same was observed as there was no classification approach finally able to create a decent model with a good AUROC.

The two models created were Logistic Regression and Gradient Boosting Classifier and hyper parameter tuning was done in both of them. However the final AUROC for LR was 71% and for GBC was 73% but as it can be seen from the dataset the imbalance of the classes is the reason behind this AUROC and not that the model is performing well.

3.4 Analysis of the results

From the cross tables there were many key takeaways such as there was no bias based on gender, race etc. while approving the loans and other correlations among the data.

However, the data did not have enough predictive power to create a classification model with action taken as the target variable.

Hence it can be concluded that the dataset is good enough for analysis purposes and the amount of interpretations one can make out of it is only limited by their imaginative power, but with this much data there cannot be a classification model possible for predicting loan action taken name.

4. Discussion and conclusion.

Decisions made: Starting with the dataset selection decision to the final model creation and data preprocessing techniques, extensive discussions, research and analysis were made. Outcome of each decision was again discussed before moving on to the next part. Some of the important decisions made were selection of dataset, handling nulls, separating categorical variables from numerical variables, models to be trained and hyper parameters to be tuned.

Difficulties faced: There were not many difficulties faced as the project was basic but in any data science project it is difficult to divide the work which is one thing where it can be said that some challenges were faced. Compared to a software project where everyone can work independently on separate components until integration is done, it is difficult to parallelly work on independent components of a data science project as many things are interdependent and hence must be performed serially instead of parallelly.

Things that worked well: All the hardwares and software components were able to handle the data properly and the process was smooth (the amount of data was not beyond the scope of machines used for this project). There were many meaningful insights while performing the data analysis part.

Things that did not work well: Mostly everything was smooth, sometimes there were minor irregularities in implementation of the project with big breaks in between. Apart from that there was a little disappointment when the models didn't perform well but that was somewhat expected from the data.

Future work identified based on your experience with this project: More data can be gathered for this project to create better models, from the banking side several models can be made to reject and accept a customer to reduce human intervention as much as possible in the loan advancement process.

Conclusion: It was a good educational project which made the basics of data mining clear whereas also helped in getting hands on experience on data preprocessing and cleaning as well as modelling techniques. There was finally no good model to predict the loan action taken but many other useful insights from the data. And anyway no results is also a result, so it can be concluded that a good model is not expected from this data.

5. Project Plan:

Who was assigned what task: Starting with dataset selection which was assigned to everyone to explore the available datasets and then after detailed discussion this project was chosen. Afterwards Ishan was assigned the data pre-processing parts such as handling nulls, binnig etc. Then Romil was assigned the data visualization part showing histograms and cross tables and finally Pallav was assigned to work on the modelling part. Even though the works were divided, each one was assigned to be actively participated in the entire project and not just their part. Finally again the presentation and the report was assigned to everyone and was done as a collaborative activity.

Who ended up doing what task: Everyone worked as per the plan mentioned above and completed their parts properly, they helped each other and were actively involved in others' parts as well.