MICROSOFT PROFESSIONAL PROGRAM IN DATA SCIENCE CAPSTONE PROJECT REPORT.

June 28, 2019

ISHAYA JEREMIAH AYOCK

Predicting Mortgage Approvals from Government Data.

Contents

0.1 Introduction

The core business of almost every financial institution is the distribution of loans. This means a significant portion of the profits that flow into these organizations come from interests that accrue on these loans. Hence one of the most principal objectives of financial institutions is to ensure that these loans are delivered into trusted hands.
Today many financial institutions approve loans after a rigorous process of verification and validation. However, cases of bad debt are still a major concern for these organizations. This situation can be improved with the employment of data science.

With available data, we can predict whether a particular applicant is a good fit for a loan by relying solely on machine learning techniques. The problem addressed is a classification problem. In a classification problem, we have to predict discrete values based on a given set of independent variables(s).

0.2 Hypothesis

Hypothesis is a very important stage in any machine learning/data science pipeline. It involves understanding the problem statement in detail by brainstorming as many factors as possible which can impact the outcome. It is always done by understanding the problem statement thoroughly before looking at the data. Below are some of the factors likely to affect the Loan Approval (these factors are the dependent variables for this loan prediction problem.)

- Applicant Income: Applicants with high income have more chances of loan approval than those with low income.
- Loan amount: Loan approval should also depend on the loan amount. If the loan amount is less, chances of loan approval should be high.
- State code: The location of the applicant could affect the approval of loan. Applicants living in states with high average incomes have higher chances of loan approval compared to those living in states with low average incomes.

0.3 Executive Summary

In this report, we present an analysis of data concerning credit scoring. This analysis is based on the Mortgage approvals data from the government. it consists of 500,000 observations of loan data with each containing specific characteristics of an individual loan applicant and the decision been made for approval.

Exploratory data analysis was carried out by calculating summary and or descriptive statistics of the numeral columns in the data set. In the Exploration of the data, several potential relationships between the loan characteristics and the accepted rate were also identified. A predictive model was built to classify a customer as eligible or not eligible for loan. Based on the exploratory data analysis, there are many factors that could affect loan approval and some of the significant features found in the given dataset are as follows.

- lender
- loan purpose
- state code
- Applicant Income
- Msa_md

0.4 Initial Data Exploration

The initial exploration of the data begins with some summary and descriptive statistics of the dataset of interest.

Individual Feature Statistics (Numeric Features)

A summary statistic for maximum, minimum, mean, median, standard deviation, and distinct count were computed for numerical columns, and the results taken from the 500,000 observations are as follows:

Column Name	Min	Max	Median	Mean	Std Dev	Count
Loan Amount	1	100878	162	221.75	590.64	500000
applicantIncome	1	10139	74	102.39	153.54	460052
ffiecmedianFamilyIncome	17858	125248	67526	69235.60	14810.06	477560
population	14	37097	4975	5416.83	2728.15	477535
minorityPopulationPct	0.53	100	22.90	31.62	26.33	477534
tractToMsaMdIncomePct	3.98	100	100	91.83	14.21	477486
numberofTo4FamilyUnits	1	13623	1753	1886.15	914.12	477435
numberOfOwnerOccupiedUnits	4	8771	1327	1427.72	737.56	477470

My Interest in this analysis is the decision to be made on whom the loan should be given to or not. Based on the training data given, out of the 500,000, 250,114(approximate 50\%) had a good loan and 249,886(approximately 50%) had a bad loan indicating that we have a balance dataset. A bar plot of the accepted column (Dependent variable) is shown as follows.

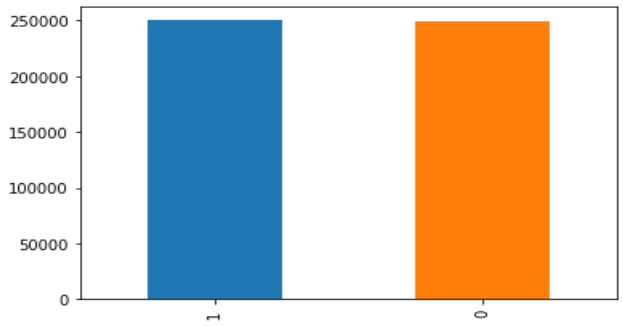

Figure 1: A bar plot of accepted column (1 means Good loan and 0 means bad loan).

In addition to the numerical values used in the table above, the following categorical features

(including levels) were present:

- preapproval: 1. Preapproval was requested, 2. Preapproval was not requested, 3. Not applicable.
- applicant ethnicity: 1. Hispanic or Latino 2. Not Hispanic or Latino, 3.Information not provided by applicant in mail, Internet, or telephone pplication, 4.Not applicable, 5.No co-applicant.

- applicant sex 1. Male, 2.Female, 3.Information not provided by applicant in mail, Internet, or telephone application, 4.Not applicable, 5.Not applicable.
- loan purpose: 1. Home purchase, 2. Home improvement, 3. Refinancing
- occupancy: 1. Owner-occupied as a principal dwelling, 2. Not owner-occupied, 3. Not applicable.
- property type: 1. One to four-family, 2. Manufactured housing, 3. Multifamily.
- loan type: 1. Conventional, 2.FHA-insured, 3.VA-guaranteed, 4.FSA/RHS.
- applicant race: 1. American Indian or Alaska Native, 2. Asian 3. Black or African American, 4.Native Hawaiian or Other Pacific Islander, 5.White, 6.Information not provided by applicant in mail, Internet, or telephone application, 7.Not applicable, 8.No co-applicant.
- msa md: A categorical with no ordering -1 indicating missing values.
- state code: A categorical with no ordering -1 indicating missing values.
- county code: A categorical with no ordering -1 indicating missing values.
- Co Applicants 1. true, 2.False
- Lender: A categorical with no ordering indicating the approving body.
- Accepted: Is an integer only valid value 0 and 1.

Based on the above features, bar charts were created to visualize the hidden information of the categorical features. And the following insight were depicted:

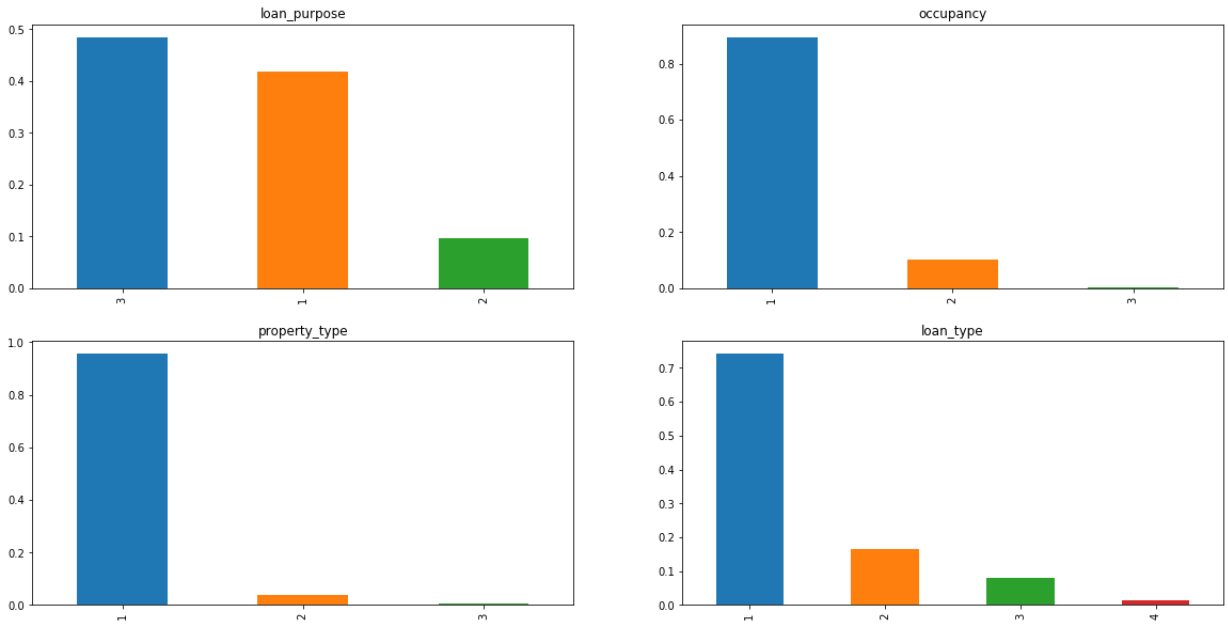

Figure 2: From the above plot, we can infer that,

1. 48% applicants went for loan purpose category 3, 42% for 1 and 10% for category 2.
2. Around 85% of the applicants are for category 1, 13% for 2 and 2% are for 3 in the occupancy field.
3. For Property type, around 95% of the applicants are for category 1, 3% for 2 and 2% are for 3.
4. For loan type, around 75% of the applicants are for category 1, 15% for 2 and 7% for 3 and 3% are for category 4.

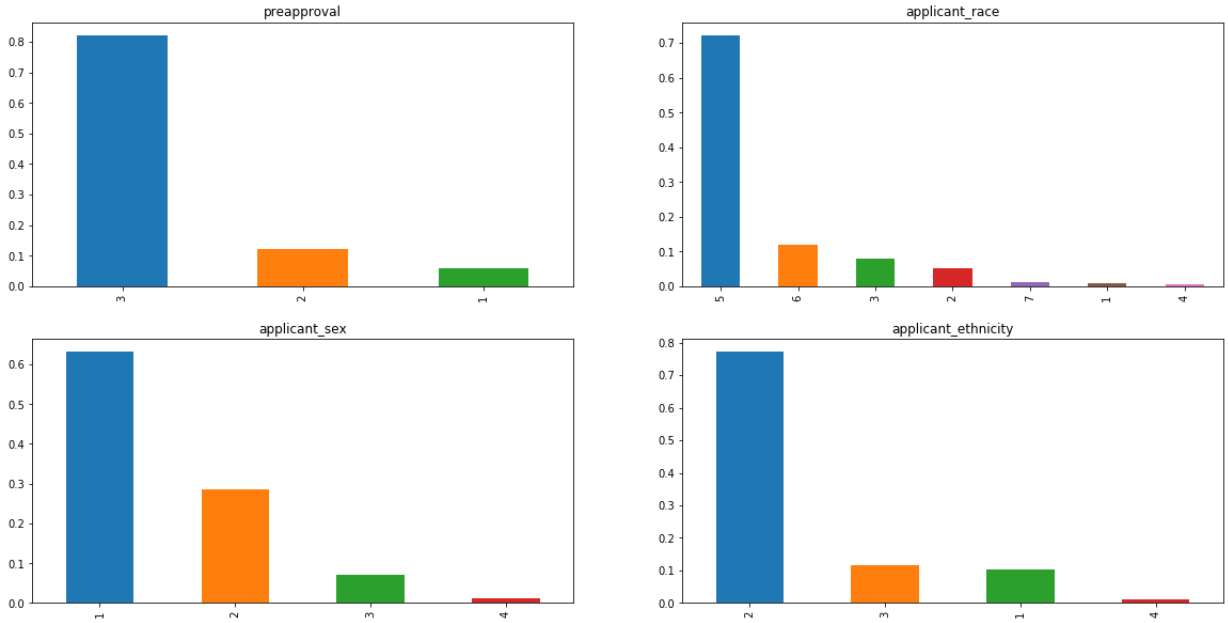

Figure 3:

1. For preapproval, 80% of the applicants are for category 3, 12% for 2 and 8%are for category 1.
2. Majority of the applicants race fall under category 5 i.e around 72%, 12%for 6, 8% for 3, 3% for 2, 4% for 7 and 1% for category 1.
3. Most of the applicant sex (65%) fall under category 1, 28% for 2, 5% for 3 and 2% for category 4.
4. For the applicant ethnicity, we can see that (78%) fall under category 2, 10%for 3, 10% for 1 and 2% for category 4.

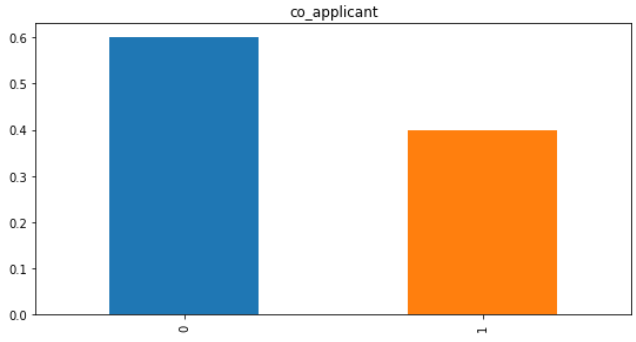

Figure 4:

Following the above bar plot, it can be inferred that most of the applicants don't have any dependents.

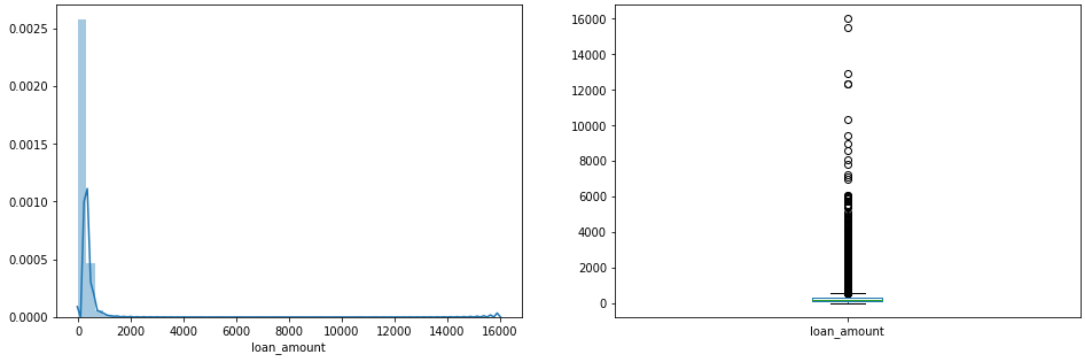

Figure 5:

It can be inferred that most of the data in the distribution of loan Amount is towards left which means it is not normally distributed. Since most algorithms works better if the data is normally distributed. Also, we can see that the box plot confirms the presence of a lot of outliers/extreme values.

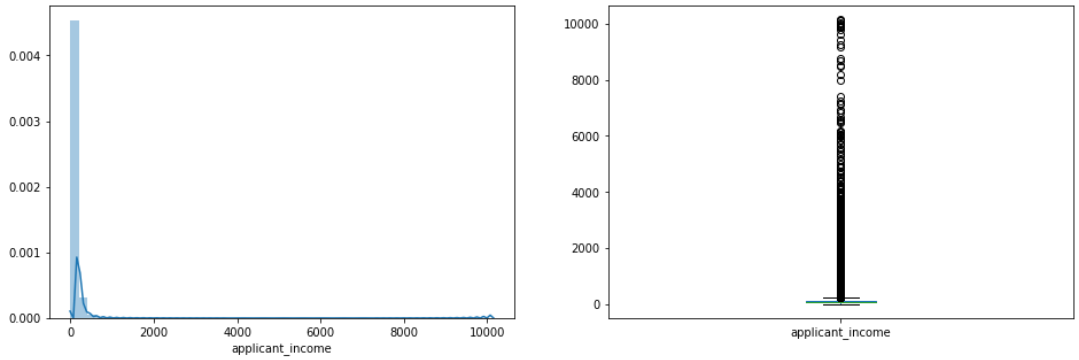

Figure 6:

Also, from the applicant income plots, we can infer that most of the data in the distribution is towards left which means it is not normally distributed. Also, we can see that the box plot confirms the presence of a lot of outliers or extreme values and there is need for treatment.

0.5 Numerical Independent Variable

After exploring the individual features, an attempt was made to identify relationships between features in the data. A scatter plot matrix was generated initially to compare numeric features with one another as shown below; The correlation between the numeric columns was then calculated with the following results:

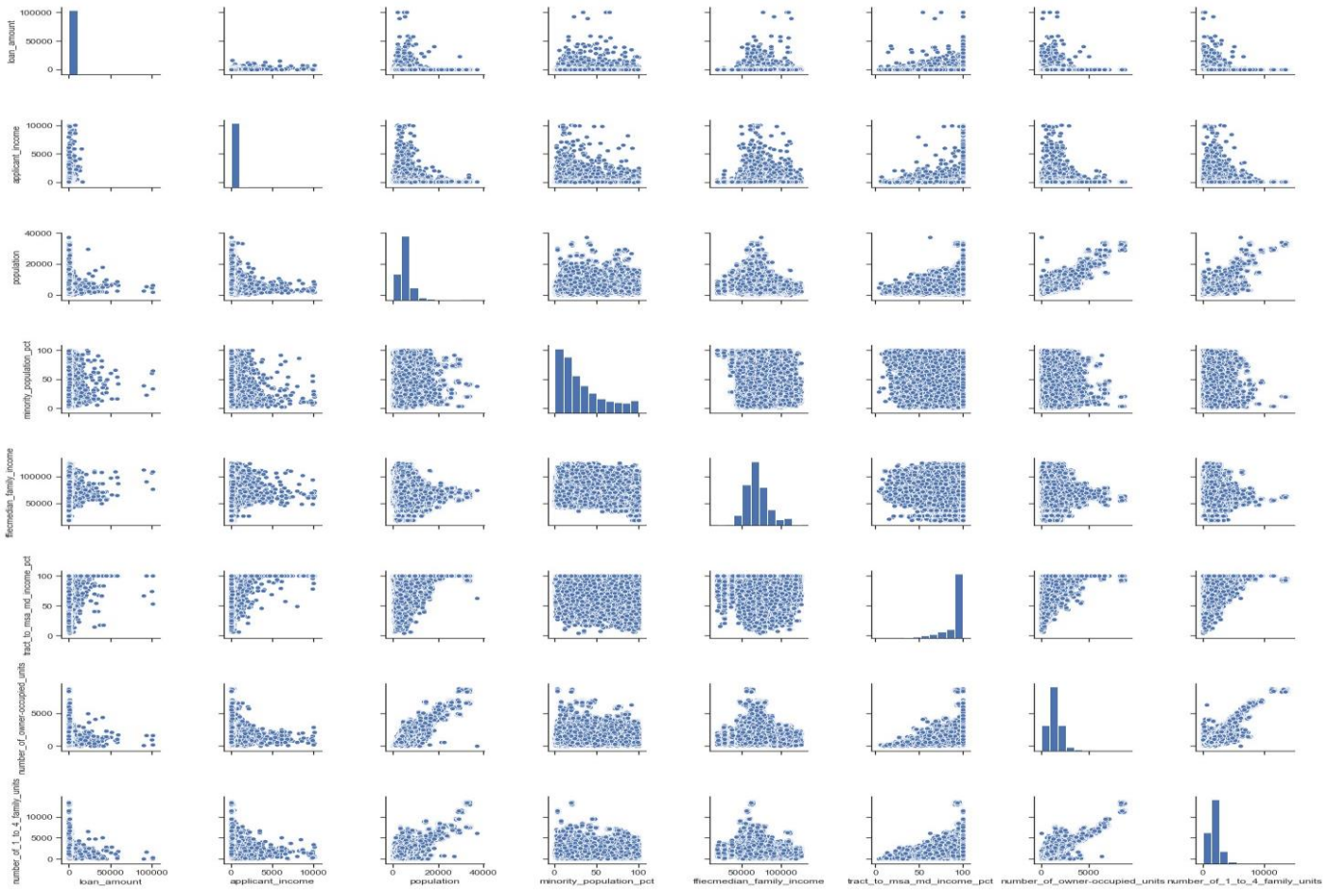

Figure 7: A pair plot of the numerical values

Correlation	loanAmnt	apIncme	ffFmIncm	poplatn	mPopulatnPct	trcIncmePct	NfOwOcupUt	NFmlyUnt
loanAmnt	**1.000**	0.006	0.080	0.004	0.010	0.012	-0.006	-0.025
apIncme	0.006	**1.000**	0.023	-0.016	-0.026	-0.013	0.005	-0.009
ffFmIncm	0.004	-0.016	**1.000**	0.307	0.705	0.704	0.392	0.307
poplatn	0.010	-0.026	0.307	**1.000**	0.452	0.454	0.885	0.856
mPopulatnPct	0.080	0.023	0.705	0.452	**1.000**	0.987	0.546	0.536
trcIncmePct	0.012	-0.013	0.704	0.454	0.987	**1.000**	0.589	0.566
NfOwOcupUt	-0.006	0.005	0.393	0.885	0.546	0.589	**1.000**	0.923
NFmlyUnt	-0.009	-0.009	0.307	0.856	0.536	0.566	0.923	**1.000**

Table 1: A table showing the correlation of the numerical features.

0.6 Categorical Independent Variable vs Target Variable

First of all, we will find the relationship between the target variable and categorical independent variables. Let us look at the stacked bar plot now which will give us the proportion of approved and the unapproved loans.

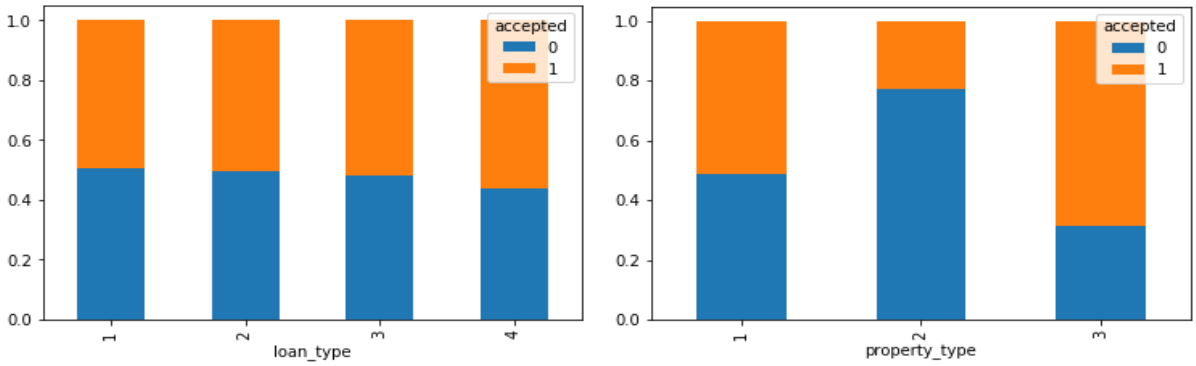

Figure 8:

1. It can be inferred that the proportion of loan type for the applicants is more or less same for both approved and unapproved loans for all categories except for that of category 4 which approved loan is 58%.
2. Likewise, for the property type, we have 75% unapproved loans in category 2 as compare to category 1 and more approved loans in category 3 with about 76%.

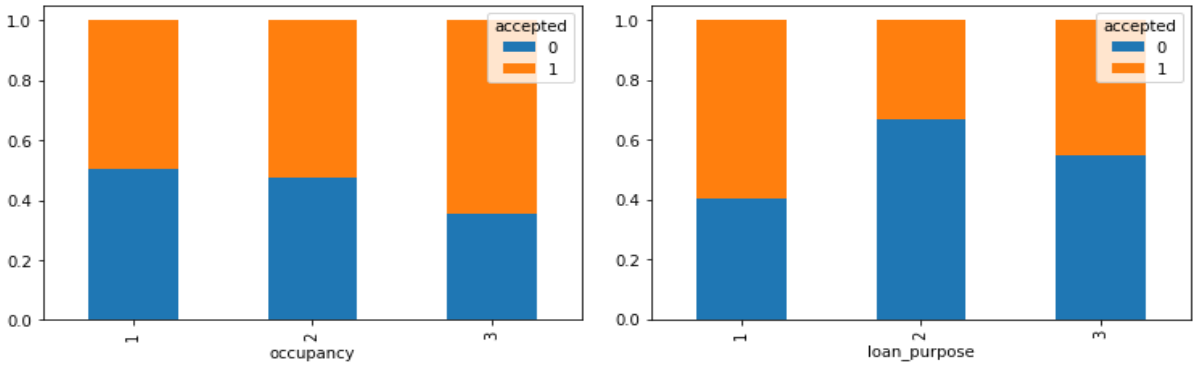

Figure 9:

1. For Occupancy, we can infer that the proportion of approved to unapproved loan is around 50% for category 1 and 2 as compare to 65% approved loans in category 3.
2. For Loan Purpose, 60% had their loan approved in category 1 as compare to class 2 and category 3.

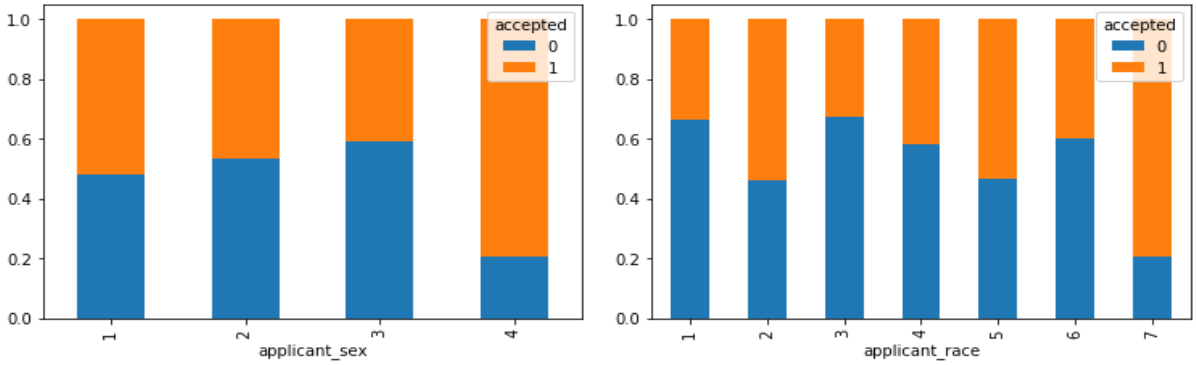

Figure 10:

1. The Proportion of loans getting approved in category 4 of applicant sex is higher as compared to that of category 2 and 3 also having a 50:50 approved loan in category 1.
2. Also, the Proportion of loan approval in category 7, 5 and 2 of the applicant race is higher as compared to that of the other categories.

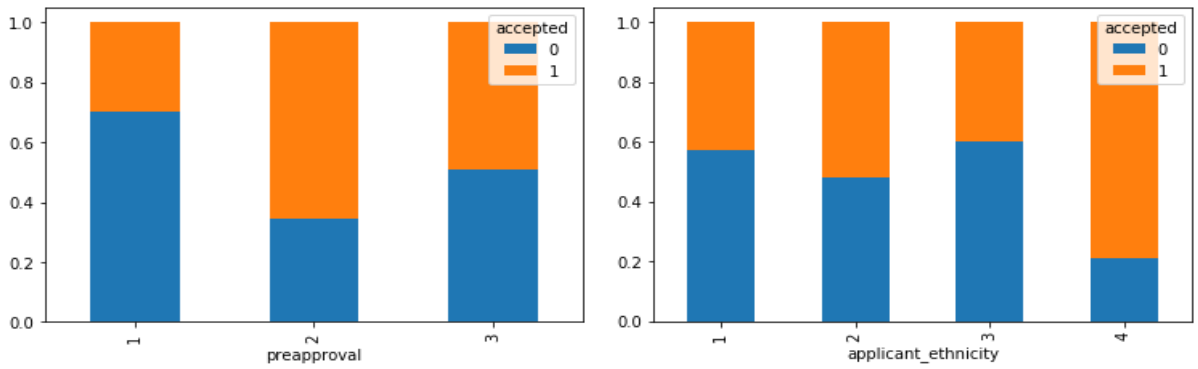

Figure 11:

1. From Preapproval, we can infer that category 2 has more loan acceptance rate as compare to category 3 and 1.
2. We can also see that category 4 of the applicant ethnicity has more loan approvals as compare to other categories.

0.7 Numerical Independent Variable vs Target Variable

We will try to find the mean values of applicants for which the loan has been approved vs the mean of applicants for which it has not been approved with data been partition into bins for easy explanations.

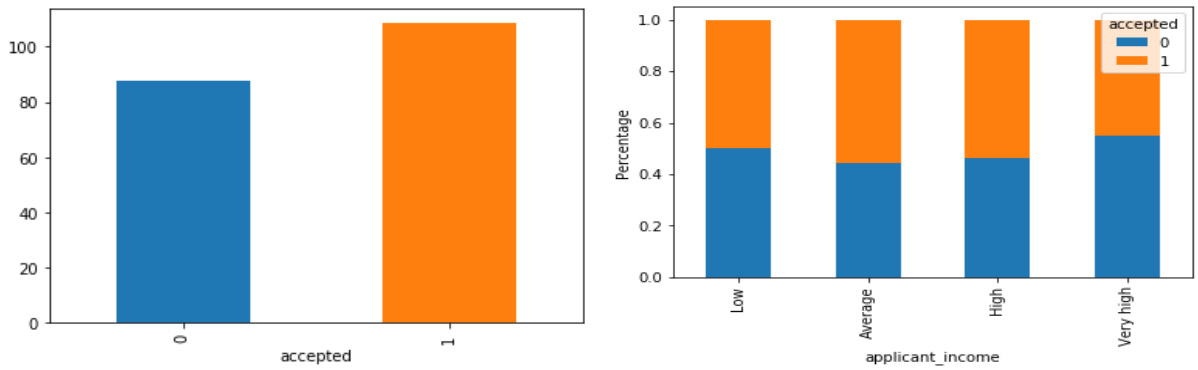

Figure 12:

1. The rate of loan acceptance is a little bit higher as compare to the unacceptance rate.

2. It can be inferred that Applicant income does not affect the chances of loan approval, which contradicts our hypothesis in which we assumed that if the applicant income is high the chances of loan approval will also be high.

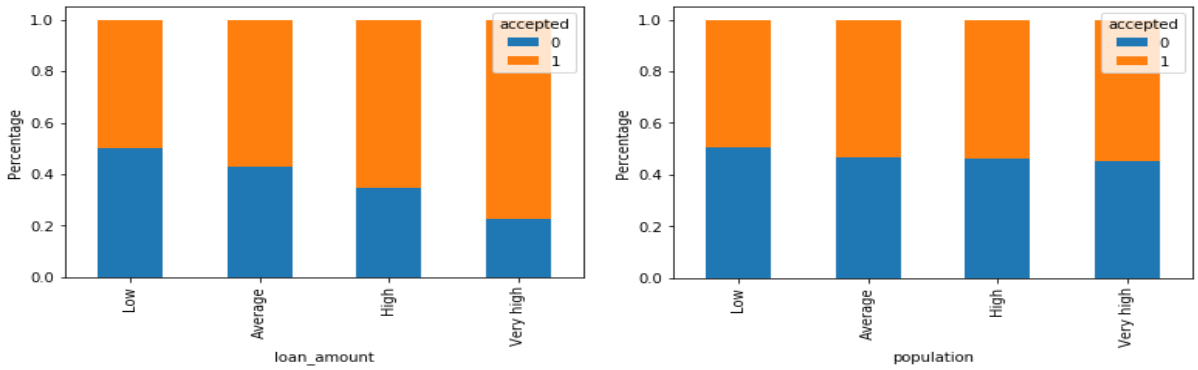

Figure 13:

1. From the Loan Amount, we can infer that applicants with higher loan amount had 80% loan approval as compare to applicant with lower loan amount which also contract our hypothesis.
2. There is nothing significant we can infer from population vs accepted plot because they all have the same percentage for both approved and unapproved loan.

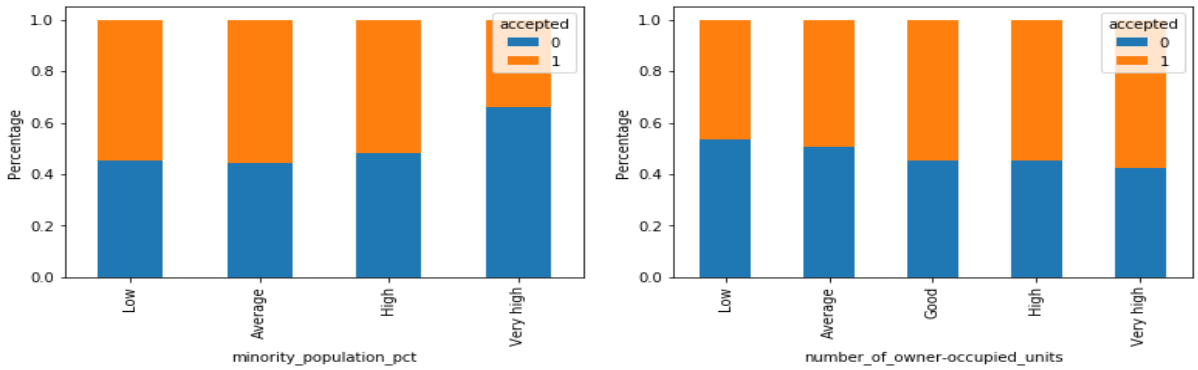

Figure 14:

1. Also, for Minority_population_pct, we can infer that applicants with a very high Minority_population_pct have less loan approval as compare to applicant with lower Minority_population_pct.

2. For No_of_owner_units, we can infer that applicants with a very highNo_of_owner_units have their loan approved more as compare to applicant with lower No_of_owner_units.
loan 1 and unapproved loan 0.

0.8 Classification of Mortgages Government data

Based on the analysis of the Mortgages government data, a predictive model was built with the features provided to classify loan applicants into two different classes, approved loan 1 and unapproved loan 0.

This model was created using two different algorithms (LightGBM and *CatBoost*) were the *CATBOOST* outperform the *LightGBM* and was choosing the model of interest. The Algorithms were trained with training set (400,000) and testing set of (100,000) and its yields the result as follows.

- True positive: 33691.
- True Negative: 39537.
- False Positive: 10629.
- False Negative: 16143.

This translates in to the following standard performance metrics for classification:

- Accuracy: 73.33%.
- Precision: 78.81%.
- Recall: 71.01%.
- F1 Score: 74.71%.

ROC

The Received Operator Characteristic (ROC) curve for the model was 0.84 and is as shown below, with the red line indicating the model's performance at varying classification threshold values, and the blue line showing the expected results of a random guess:

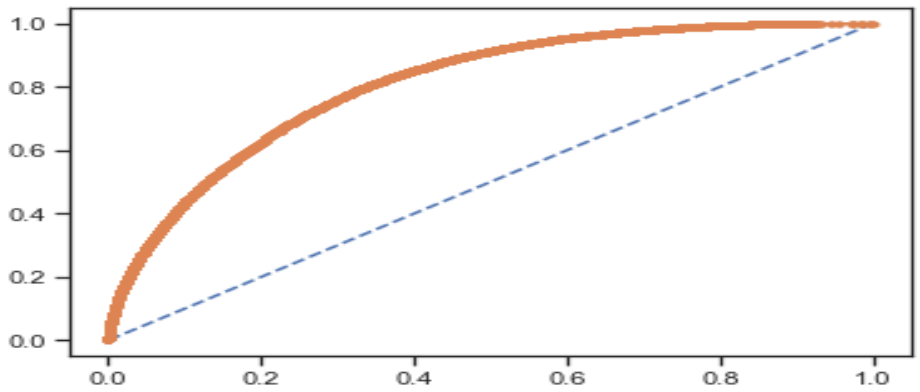

Figure 15: ROC plot

0.9 Conclusion

In this analysis, we can see that the system can predict whether a particular loan applicant is safe or not safe for giving out loan based on his/her characteristics. And the whole process of validation is automated using machine learning techniques.

In particular, the features such as lender, loan purpose, state code, applicant's income, and Msa_md have a significant effect on the loan acceptance.