NEW YORK HMDA DATASET.

LOAN STATUS PREDICTION

Mentored by: Ms. Vidya K

Submitted by:

5. Sanjana Basu
6. Ajithkumar Velmuruga
7. Mukka Rakesh
8. Sharma R

Contents

Abstract: 2
Introduction: 2
Solution Offered: 2
The Domain and Dataset: 2
Flow of Geographic Data: 9
Pre-Processing Data Analysis: 10
Dataset observations: 10
Null/Missing Value Treatment: 10
Redundant columns: 11
Single value columns: 11
Other columns: 11
Feature Profiling: 11
Null values Imputation: 12
Project Justification: 13
Data Exploration (EDA): 13
Categorical 13
Target- Action_taken: 13
Purchaser Type: 14
Conforming_loan_limit: 14
Loan_type: 15
Loan_purpose: 15
Lien_status, reverse_mortgage, open-end_line_of_credit, business_or_commercial_purpose, loan_to_value_ratio: 16
Property_value, interest_only_payment, occupancy type, debt_to_income_ratio: 17
Applicant Age, sex, ethnicity and race: 18
Reason for denial, family dwelling type: 19
B. Numerical: 20
Check the Distribution: 20
Numeric vs Target: 20
Multicollinearity: 21
Outlier Detection: 23
Statisticalsignificanceofvariables: 24
Check Class Imbalance: 26
APPROACH AND STEPS TAKEN TO SOLVE THE BUSINESS PROBLEM: 26
Step 1: Business Understanding 26

Step 2: Data Understanding 27
Step 3: Data Preparation 27
STEP 4: Modeling 27
STEP 5: Evaluation 28
MODEL BUILDING: 28
Removing some more columns: 28
Feature Encoding: 28
Feature Scaling: 29
Base Model: 29
Other Models: 30
Feature Selection: 30
Cross validating models: 31
MODEL EVALUATION: 32
Confusion Matrix: 32
Classification Report: 33
Roc_auc score: 33
Implications: 34
Limitations: 35
Conclusion: 35

Abstract:

The Home Mortgage Disclosure Act (HMDA) was enacted by Congress in 1975 and was implemented by the Federal Reserve Board's Regulation C. This regulation provides the public loan data that can be used to assist the public officials in distributing public-sector investments so as to attract private investment to areas where it is needed and in identifying possible discriminatory lending patterns. Nowadays, there are numerous risks related to providing loans both for the financial institutions and the borrowers getting the loans. Financial Institutions need to analyze their customers for loan eligibility so that they can also target those customers. These Institutions may want to automate the loan eligibility process (real time) based on customer details, since the volume of data is rapidly growing day-by-day. This data can be used to analyze the customer's behavior and the risk around loan can be reduced. Also these Institutions receive numerous loan requests each day and it may take a while to process a particular application and make a decision regarding its eligibility for loan.

Introduction:

The **Home Mortgage Disclosure Act (HMDA**) requires many financial institutions to maintain, report, and publicly disclose loan-level information about mortgages**.**This data helps in showing whether lenders are serving the housing needs of their communities; they provide public officials information that helps them in making decisions and policies, and they shed light on lending patterns that could be discriminatory. The public data is modified to protect applicant and borrower privacy. **The goal is to create greater transparency and to protect borrowers in the residential mortgage market and monitor the geographic targets of mortgage lenders, providing a way to identify predatory or discriminatory lending practices.**

Solution Offered:

Automating the loan approval process can save a huge amount of time and man-power for a Financial Institution. Machine Learning can be used to automate the loan eligibility process and provide near-accurate predictions on whether a loan application should be approved or not based on the customer's financial information and demographics. We are trying to build a model which takes around 90 identifiers like the sex, race, and income of those applying for obtaining mortgages and predict whether the loan is approved or not. This would automate the whole process of predicting the approval of loan for an individual based on customer's identifications.

The Domain and Dataset:

HMDA data can be used to identify indicators of potential mortgage discrimination, however HMDA does not contain sufficient data to make conclusive determinations regarding discrimination. It is important to understand that in all cases of possible discrimination, the basic regulatory inquiry revolves around whether a protected class of persons being denied a loan or offered different terms for reasons other than objectively acceptable characteristics (e.g. income, collateral). This data also allows regulators, public officials and consumer watchdogs to monitor trends in mortgage borrowing and lending for compliance with fair housing and other laws and to direct housing investment and government funding to areas where it is needed. As a result,regulatory authorities can evaluate whether the lender is adequately meeting the needs of the prospective borrowers in that community.Here in this project we are using the below identifications in the records to determine the action taken whether the loan is approved or rejected.

FEATURE NAME	FEATURE DESCRIPTION	DATA TYPE	MISSING VALUE %
activity_year	The calendar year the data submission covers	Numeric	0
lei	A financial institution's Legal Entity Identifier	Alphanumeric	0
derived_msa_md	The 5 digit derived MSA (metropolitan statistical area) or MD (metropolitan division) code. An MSA/MD is an area that has at least one urbanized area of 50,000 or more population.	Alphanumeric	0
state_code	Two-letter state code	Alphanumeric	0
county_code	State-county FIPS code	Alphanumeric	0 .61
census_tract	11 digit census tract number	Alphanumeric	0.74
conforming_loan_limit	Derived loan product type from Loan Type and Lien Status fields for easier querying of specific records	Alphanumeric	0.23
derived_loan_product_type	Derived loan product type from Loan Type and Lien Status fields for easier querying of specific records	Alphanumeric	0
derived_dwelling_category	Derived dwelling type from Construction Method and Total Units fields for easier querying of specific records	Alphanumeric	0
derived_ethnicity	Indicates whether the reported loan amount exceeds the GSE (government sponsored enterprise) conforming loan limit	Alphanumeric	0
derived_race	Single aggregated race categorization derived from applicant/borrower and co-applicant/co-borrower race fields	Alphanumeric	0
derived_sex	Single aggregated sex categorization derived from applicant/borrower and co-applicant/co-borrower sex fields	Alphanumeric	0
action_taken	The action taken on the covered loan or application	Numeric	0
purchaser_type	Type of entity purchasing a covered loan from the institution	Numeric	0
preapproval	Whether the covered loan or application involved a request for a preapproval of a home purchase loan under a	Numeric	0

	preapproval program		
loan_type	The type of covered loan or application	Numeric	0
loan_purpose	The purpose of covered loan or application	Numeric	0
lien_status	Lien status of the property securing the covered loan, or in the case of an application, proposed to secure the covered loan	Numeric	0
reverse_mortgage	Whether the covered loan or application is for a reverse mortgage	Numeric	0
open_end_line_of_credit	Whether the covered loan or application is for an open-end line of credit	Numeric	0
business_or_commercial_purpose	Whether the covered loan or application is primarily for a business or commercial purpose.	Numeric	0
loan_amount	The amount of the covered loan, or the amount applied for	Alphanumeric	0
combined_loan_to_value_ratio	The ratio of the total amount of debt secured by the property to the value of the property relied on in making the credit decision	Alphanumeric	46.02
interest_rate	The interest rate for the covered loan or application	Alphanumeric	58.07
rate_spread	The difference between the covered loan's annual percentage rate (APR) and the average prime offer rate (APOR) for a comparable transaction as of the date the interest rate is set.	Alphanumeric	98.73
hoepa_status	Whether the covered loan is a high-cost mortgage	Numeric	0
total_loan_costs	The amount, in dollars, of total loan costs	Alphanumeric	69.34
total_points_and_fees	The total points and fees, in dollars, charged in connection with the covered loan	Alphanumeric	98.73
origination_charges	The total of all itemized amounts, in dollars, that are designated borrower-paid at or before closing	Alphanumeric	67.81

discount_points	The points paid, in dollars, to the creditor to reduce the interest rate	Alphanumeric	87.68
lender_credits	The amount, in dollars, of lender credits	Alphanumeric	89.47
loan_term	The number of months after which the legal obligation will mature or terminate, or would have matured or terminated.	Alphanumeric	2.14
prepayment_penalty_term	The term, in months, of any prepayment penalty	Alphanumeric	92.79
intro_rate_period	The number of months, or proposed number of months in the case of an application, until the first date the interest rate may change after closing or account opening	Alphanumeric	79.42
negative_amortization	Whether the contractual terms include, or would have included, a term that would cause the covered loan to be a negative amortization loan	Numeric	0
interest_only_payment	Whether the contractual terms include, or would have included, interest-only payments	Numeric	0
balloon_payment	Whether the contractual terms include, or would have included, a balloon payment	Numeric	0
other_nonamortizing_features	Whether the contractual terms include, or would have included, any term, other than those described in Paragraphs 1003.4(a)(27)(i), (ii), and (iii) that would allow for payments other than fully amortizing payments during the loan term	Numeric	0
property_value	The value of the property securing the covered loan or, in the case of an application, proposed to secure the covered loan, relied on in making the credit decision	Alphanumeric	4.80
construction_method	Construction method for the dwelling	Numeric	0
occupancy_type	Occupancy type for the dwelling	Numeric	0
manufactured_home_secure	Whether the covered loan or application is, or would have been, secured by a manufactured home and land, or by a	Numeric	0

d_property_type	manufactured home and not land		
manufactured_home_land_property_interest	The applicant's or borrower's land property interest in the land on which a manufactured home is, or will be, located	Numeric	0
total_units	The number of individual dwelling units related to the property securing the covered loan or, in the case of an application, proposed to secure the covered loan	Alphanumeric	0
multifamily_affordable_units	Reported values as a percentage, rounded to the nearest whole number, of the value reported for Total Units	Alphanumeric	98.67
income	The gross annual income, in thousands of dollars, relied on in making the credit decision, or if a credit decision was not made, the gross annual income relied on in processing the application	Alphanumeric	26.08
debt_to_income_ratio	The ratio, as a percentage, of the applicant's or borrower's total monthly debt to the total monthly income relied on in making the credit decision	Alphanumeric	45.42
applicant_credit_score_type	The name and version of the credit scoring model used to generate the credit score, or scores, relied on in making the credit decision	Numeric	0
co_applicant_credit_score_type	The name and version of the credit scoring model used to generate the credit score, or scores, relied on in making the credit decision	Numeric	0
applicant_ethnicity_1	Ethnicity of the applicant or borrower	Alphanumeric	0.10
applicant_ethnicity_2	Ethnicity of the applicant or borrower	Alphanumeric	97.61
applicant_ethnicity_3	Ethnicity of the applicant or borrower	Alphanumeric	99.95
applicant_ethnicity_4	Ethnicity of the applicant or borrower	Alphanumeric	99.99
applicant_ethnicity_5	Ethnicity of the applicant or borrower	Alphanumeric	99.99
co_applicant_ethnicity_1	Ethnicity of the first co-applicant or co-borrower	Alphanumeric	0.02
co_applicant_ethnicity_2	Ethnicity of the first co-applicant or co-borrower	Alphanumeric	99.26
co_applicant_ethnicity_3	Ethnicity of the first co-applicant or co-borrower	Alphanumeric	99.98

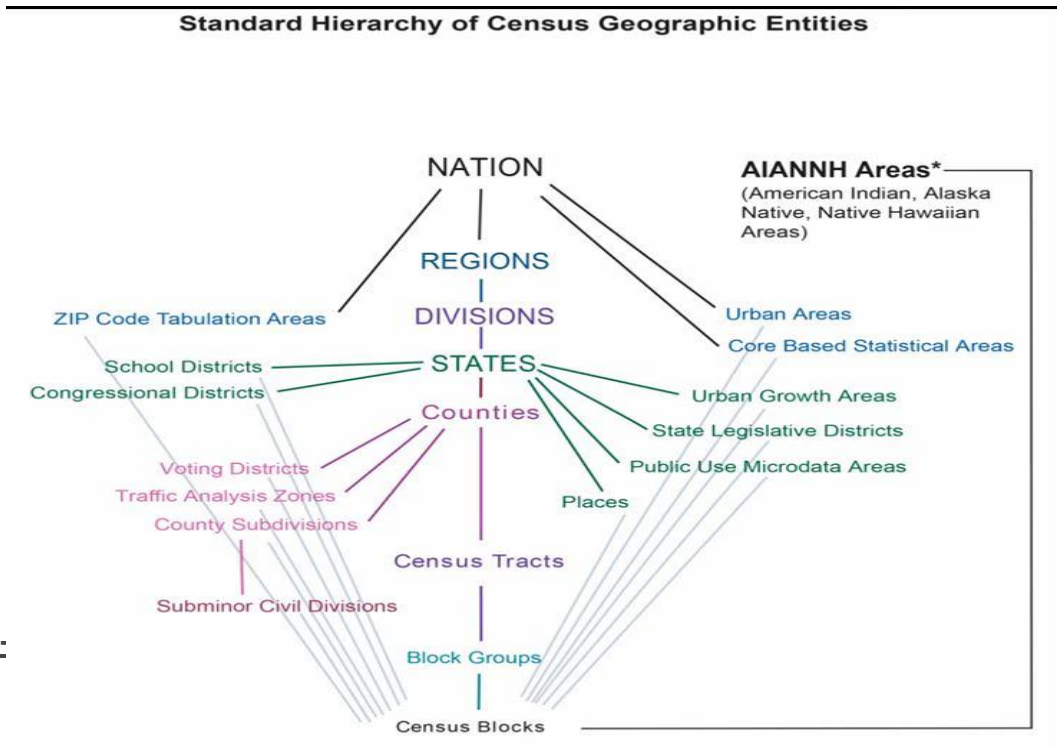
co_applicant_ethnicity_4	Ethnicity of the first co-applicant or co-borrower	Alphanumeric	99.99
co_applicant_ethnicity_5	Ethnicity of the first co-applicant or co-borrower	Alphanumeric	99.99
applicant_ethnicity_observed	Whether the ethnicity of the applicant or borrower was collected on the basis of visual observation or surname	Alphanumeric	0
co_applicant_ethnicity_observed	Whether the ethnicity of the first co-applicant or co-borrower was collected on the basis of visual observation or surname	Alphanumeric	0
applicant_race_1	Race of the applicant or borrower	Alphanumeric	0.03
applicant_race_2	Race of the applicant or borrower	Alphanumeric	96.22
applicant_race_3	Race of the applicant or borrower	Alphanumeric	99.79
applicant_race_4	Race of the applicant or borrower	Alphanumeric	99.98
applicant_race_5	Race of the applicant or borrower	Alphanumeric	99.99
co_applicant_race_1	Race of the first co-applicant or co-borrower	Alphanumeric	0.006
co_applicant_race_2	Race of the first co-applicant or co-borrower	Alphanumeric	98.68
co_applicant_race_3	Race of the first co-applicant or co-borrower	Alphanumeric	99.93
co_applicant_race_4	Race of the first co-applicant or co-borrower	Alphanumeric	99.99
co_applicant_race_5	Race of the first co-applicant or co-borrower	Alphanumeric	99.99
applicant_race_observed	Whether the race of the applicant or borrower was collected on the basis of visual observation or surname	Numeric	0
co_applicant_race_observed	Whether the race of the first co-applicant or co-borrower was collected on the basis of visual observation or surname	Numeric	0
applicant_sex	Sex of the applicant or borrower	Numeric	0

co_applicant_sex	Sex of the first co-applicant or co-borrower	Numeric	0
applicant_sex_observed	Whether the sex of the applicant or borrower was collected on the basis of visual observation or surname	Numeric	0
co_applicant_sex_observed	Whether the sex of the first co-applicant or co-borrower was collected on the basis of visual observation or surname	Numeric	0
applicant_age	The age, in years, of the applicant or borrower	Alphanumeric	0
co_applicant_age	The age, in years, of the first co-applicant or co-borrower	Alphanumeric	0
applicant_age_above_62	Whether the applicant or borrower age is above 62	Alphanumeric	28.63
co_applicant_age_above_62	Whether the first co-applicant or co-borrower age is above 62	Alphanumeric	73.44
submission_of_application	Whether the applicant or borrower submitted the application directly to the financial institution	Alphanumeric	0
initially_payable_to_institution	Whether the obligation arising from the covered loan was, or, in the case of an application, would have been, initially payable to the financial institution	Alphanumeric	0
aus_1	The automated underwriting system(s) (AUS) used by the financial institution to evaluate the application	Numeric	0
aus_2	The automated underwriting system(s) (AUS) used by the financial institution to evaluate the application	Alphanumeric	95.26
aus_3	The automated underwriting system(s) (AUS) used by the financial institution to evaluate the application	Alphanumeric	97.39
aus_4	The automated underwriting system(s) (AUS) used by the financial institution to evaluate the application	Alphanumeric	99.97
aus_5	The automated underwriting system(s) (AUS) used by the financial institution to evaluate the application	Alphanumeric	99.98
denial_reason_1	The principal reason, or reasons, for denial	Alphanumeric	0
denial_reason_2	The principal reason, or reasons, for denial	Alphanumeric	87.41
denial_reason_3	The principal reason, or reasons, for denial	Alphanumeric	97.75

denial_reason_4	The principal reason, or reasons, for denial	Alphanumeric	99.75
tract_population	Total population in tract	Numeric	0
tract_minority_population_percent	Percentage of minority population to total population for tract, rounded to two decimal places	Numeric	0
ffiec_msa_md_median_family_income	FFIEC Median family income in dollars for the MSA/MD in which the tract is located (adjusted annually by FFIEC)	Numeric	0
tract_to_msa_income_percentage	Percentage of tract median family income compared to MSA/MD median family income	Numeric	0
tract_owner_occupied_units	Number of dwellings, including individual condominiums, that are lived in by the owner	Numeric	0
tract_one_to_four_family_homes	Dwellings that are built to houses with fewer than 5 families	Numeric	0
tract_median_age_of_housing_units	Tract median age of homes	Numeric	0

Flow of Geographic Data:

The dataset consists of three location based fields(geographic entities) like derived_msa,county_code and cesus_tract,understanding their order of hierarchy is significant.The below flow chart represents the hierarchy of Geographic entities:



Pre-

As real-world financial data are often clumsy, a particular focus will be led on pre-processing task handling missing data, duplicates, redundant columns, outliers and data pollution thereby able to reduce the dataset features and optimizing it for further before model
deployment.

Dataset observations:

1. There are a total of 99 columns in the raw data and the number of data rows are 182272.
2. There are over 60+ columns that are categorical but nominal in nature such as race, ethnicity etc.
3. These nominal values need to be converted to their respective categories,otherwise deriving inference can be tricky.
4. Columns relating to census tracts will have 0 value present because there can be areas with no population or activity.
5. The dataset consist of redundant columns mostly named as 'derived' columns which are generalized form of the original columns.These columns must be removed.

Null/Missing Value Treatment:

- The first step in pre-processing the data consist of handling missing values. Missing values refers to the absence, voluntary or not, of data in a record. While the initial step is to identify and encode missing values, the second step consists in addressing the missing values. Each variable comprising missing values were independently analysed, and depending upon the percentage of the missing values we either drop the column or impute it with values using some statistical criteria.
- As a general rule, variables with 50% or more missing values should be dropped from the analysis and moreover we don't find any features helping in imputing these missing values. So, in our dataset,out of 99 columns 34 columns have more than 50% of missing data,hence we can drop those columns for further analysis.

Redundant columns:

Now we are left with 65 columns.Among these columns there are redundant columns that have been modified from the original loan record to protect privacy, or added to enhance the usefulness of the data. Derived fields were aggregated from the source data, following a methodology developed by the CFPB for analysis and privacy protection, but are not an official and sole government definition.Some of these columns are more generalized which may reduce the scope of EDA.Hence this kind of columns should be dropped.

```
df['applicant_race-1'].nunique(),df['derived_race'].nunique()
```

```
(18, 9)
```

Single value columns:

The columns activity_year,state_code and preapproval have only one unique value.Since there is no variability in these columns, they should be dropped.

```
#dropping derived/unique value columns
df=df.drop(['activity_year','state_code','preapproval','derived_race','derived_ethnicity','derived_loan_product_type',
            'derived_sex','co-applicant_ethnicity_observed','applicant_ethnicity_observed','applicant_race_observed',
            'co-applicant_race_observed','applicant_sex_observed','co-applicant_sex_observed'],axis=1)
```

Other columns:

The below categorical columns are dropped since one of the categories is very highly dominating the distribution than the other categories.

```
df['manufactured_home_secured_property_type'].value_counts(normalize=True)*100
```

```
3       96.62
2        1.35
1111     1.24
1        0.80
Name: manufactured_home_secured_property_type, dtype: float64
```

```
df=df.drop(['manufactured_home_secured_property_type','manufactured_home_land_property_interest'],axis=1)
df.shape
```

```
(182272, 48)
```

Finally we are left with 48 columns,after removing all the unnecessary columns.

Feature Profiling:

- All the columns with nominal values are replaced with their respective categorical values,since it will be tricky to draw inference from nominal values while performing EDA.

```
df['action_taken']=df['action_taken'].replace({3:'denied',6:'approved'})
df['purchaser_type']=df['purchaser_type'].replace({0:'NA',1:'Fannie Mae',2:'Ginnie Mae',3:'Freddie Mac',
                                                   4:'Farmer Mac',5:'Private securitizer',6:'Commercial bank',
                                                   71:'finance company',72:'Life insurance company',8:'Affiliate institution',
                                                   9:'Others'})
df['loan_type']=df['loan_type'].replace({1:'Conventional',2:'FHA',3:'VA',4:'RHS'})
df['loan_purpose']=df['loan_purpose'].replace({1:'Home purchase',2:'Home improvement',31:'Refinancing',32:'Cash-out refinancing'
                                               4:'Other purpose',5:'NA'})
df['lien_status']=df['lien_status'].replace({1:'FL',2:'SL'})
df['reverse_mortgage']=df['reverse_mortgage'].replace({1:'Yes',2:'No',1111:'Exempt'})
df['open-end_line_of_credit']=df['open-end_line_of_credit'].replace({1:'Yes',2:'No',1111:'Exempt'})
df['business_or_commercial_purpose']=df['business_or_commercial_purpose'].replace({1:'Yes',2:'No',1111:'Exempt'})
```

- Although the columns debt_to_income_ratio,loan_to_value_ratio,property_value are numeric in nature, the inclusion of 'Exempt'(free from an obligation) value has forced them as categorical variables.Hence,these columns are binned using pd.cut() method.

Before binning:

```
df['loan_to_value_ratio'].value_counts()
```

```
80.0      4239
Exempt    2310
80.0      1794
75.0      1558
96.5      1531
          ...
95.074       1
85.686       1
57.516       1
54.187       1
95.463       1
Name: loan_to_value_ratio, Length: 35213, dtype: int64
```

After binning:

```
df['loan_to_value_ratio'].value_counts()
```

```
75-100%    36721
50-75%     31808
25-50%     15920
0-25%       7170
>100%       4456
Exempt      2310
Name: loan_to_value_ratio, dtype: int64
```

- The column total_units have same value in both number and string format,hence these values are binned as '<5 '.

```
df['total_units'].value_counts()
```

```
1          122264
1           36149
2           14827
2            3964
3            2842
4             955
3             629
5-24          294
4             218
25-49          69
50-99          35
>149           18
100-149         8
Name: total_units, dtype: int64
```

```
df['total_units'].value_counts()
```

```
<5         181848
5-24          294
25-49          69
50-99          35
>149           18
100-149         8
Name: total_units, dtype: int64
```

Null values Imputation:

There are thirteen columns with null values and need to be treated based on their datatype.For a categorical feature we have imputed with mode() value of the feature, and for numeric column we have imputed with median() value of the column.

```
county_code                         0.61
census_tract                        0.74
conforming_loan_limit               0.23
loan_to_value_ratio                46.02
loan_term                           2.14
property_value                      4.86
income                             26.09
debt_to_income_ratio               45.42
applicant_ethnicity-1               0.11
co-applicant_ethnicity-1            0.02
applicant_race-1                    0.04
co-applicant_race-1                 0.01
ffiec_msa_md_median_family_income   0.61
dtype: float64
```

Project Justification:

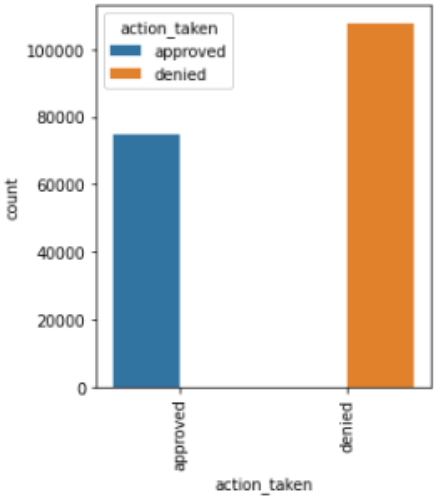
➢ The dataset that we are going to use is the real-time dataset which was collected from Home Mortage Disclosure Act(HMDA) using the API provided by the website.
➢ The dataset consist of details of borrowers who were applying for mortgages in the state of New York for the year 2020.
➢ This is a BinaryClassification problem. The dependent variable is action_taken which represents whether a particular loan request is approved or not.
➢ We can use Classification model algorithms like Logistic Regression, KNN, Decision Tree, SVM,Random Forest, etc., We can use bagging and boosting techniques to further increase the accuracy and performance of the model.

Data Exploration (EDA):

A. Categorical

1. Target- Action_taken:

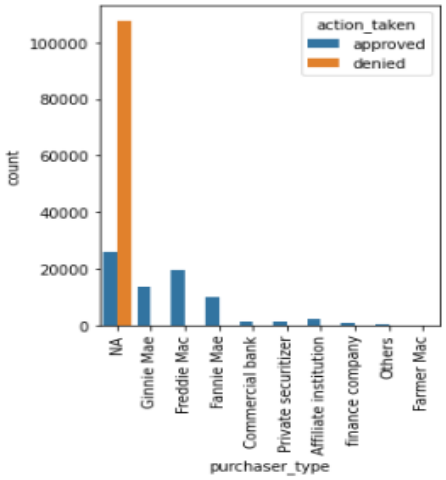
```
denied      59.05
approved    40.95
Name: action_taken, dtype: float64
NUMBER OF CATEGORIES= 2
```



★ 59% of the target variable is of class ‘denied’ and 40.95% is class ‘approved’.

2. Purchaser Type:

```
NA                      73.17
Freddie Mac             10.63
Ginnie Mae               7.44
Fannie Mae               5.54
Affiliate institution    1.22
Commercial bank          0.72
Private securitizer      0.70
finance company          0.36
Others                   0.23
Farmer Mac               0.00
Name: purchaser_type, dtype: float64
NUMBER OF CATEGORIES= 10
```

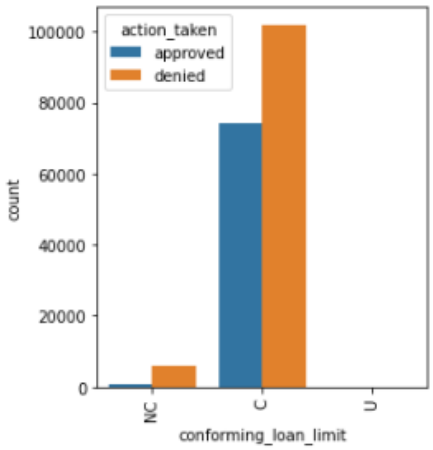


★ Except for the customers for whom purchaser type is not applicable all the other customers with any of the purchaser type are getting loan approval.

★ 73% of all customers do not have a purchaser type.

3. Conforming_loan_limit:

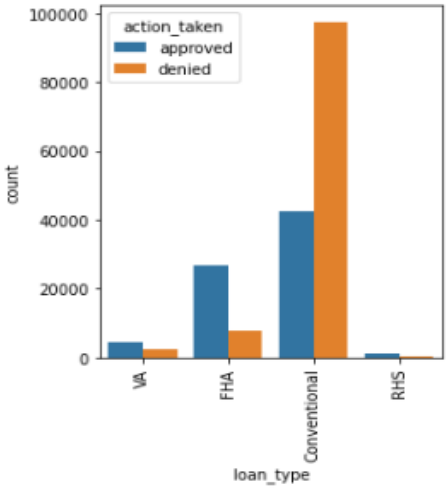
```
C     96.46
NC     3.50
U      0.04
Name: conforming_loan_limit, dtype: float64
NUMBER OF CATEGORIES= 3
```



- ★ Mortgages with Non-confirming loan limits(NC) are getting more rejection than Conforming(C).
- ★ 96% of the customers come with confirming loan limits.

4. **Loan_type:**

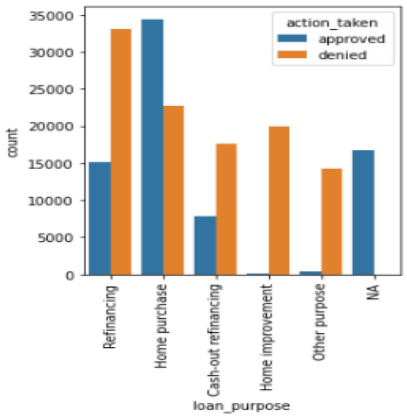
```
Conventional   76.66
FHA            18.92
VA              3.77
RHS             0.64
Name: loan_type, dtype: float64
NUMBER OF CATEGORIES= 4
```



- ★ Loans are provided more frequently for purchasing homes or for the borrowers for whom the purpose of the loan is not necessary.
- ★ 76% of all customers take loans that are not guaranteed by the FHA. This shows people prefer the conventional path of loan.

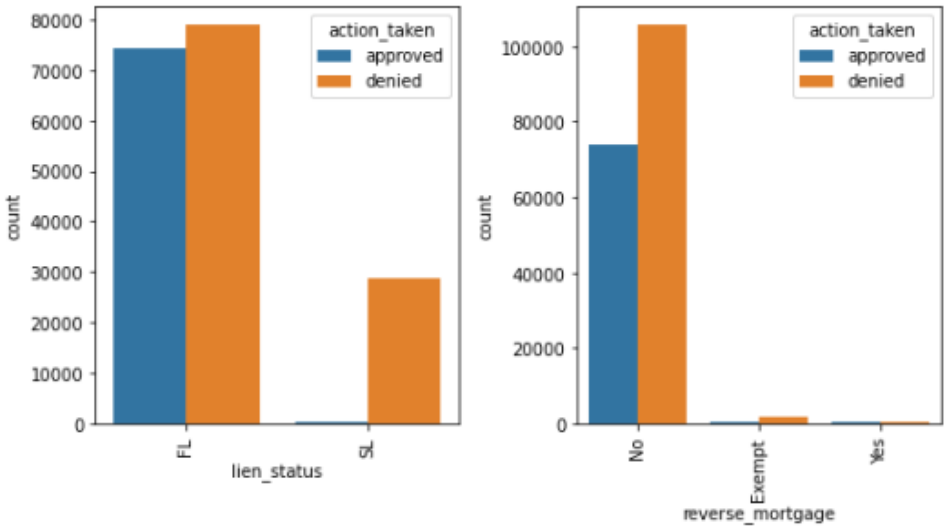
5. **Loan_purpose:**

```
Home purchase           31.32
Refinancing             26.43
Cash-out refinancing    13.99
Home improvement        11.01
NA                       9.19
Other purpose            8.06
Name: loan_purpose, dtype: float64
NUMBER OF CATEGORIES= 6
```

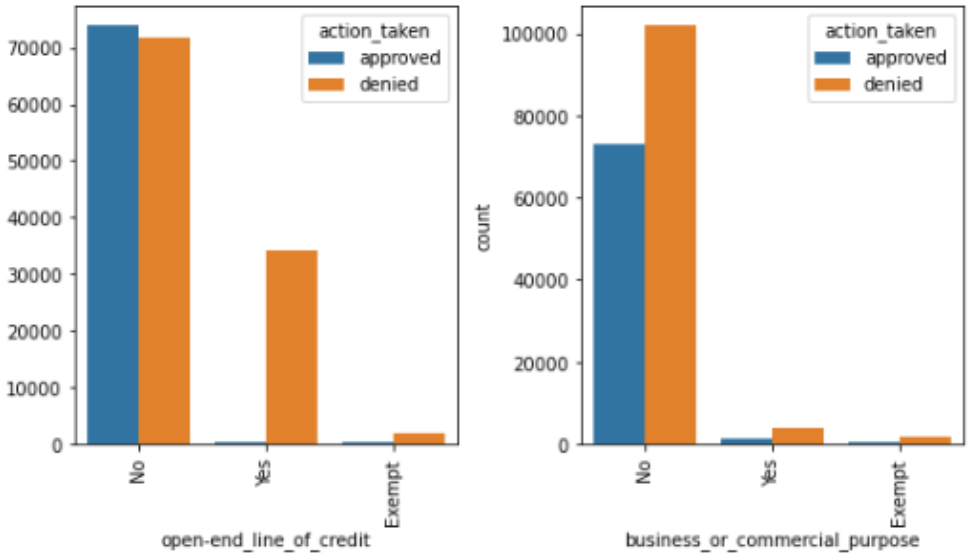


- ★ Borrowers are most likely to get loan rejection, if they are applying for home improvement or if the purpose of the loan is not known(other purpose) properly .
- ★ Most customers carry out a home loan to buy a home. A similar amount of people too carry out loans for refinancing.
- ★ 9% of the people straight up get a loan approved for reasons that need not be given.

6. **Lien_status, reverse_mortgage, open-end_line_of_credit, business_or_commercial_purpose, loan_to_value_ratio:**

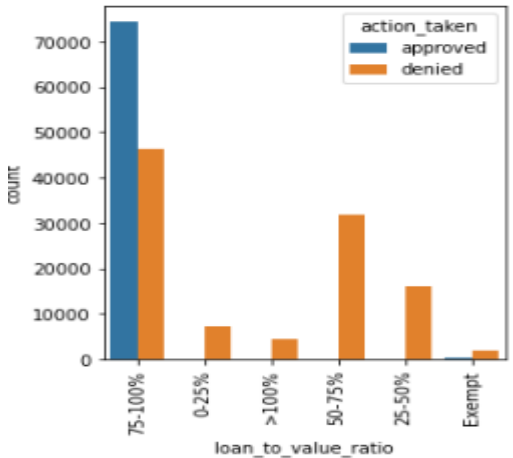


★ Loans are mostly getting rejected for a subordinate lien secured loan.
★ Most of the non reverse mortgaged cases were rejected showing that reverse mortgage is in fact an important parameter.



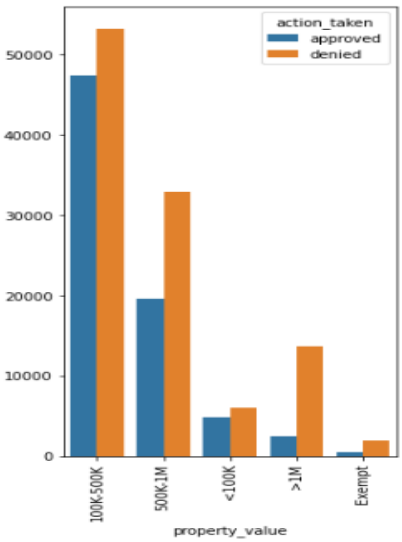
★ Most loans requests in the data are not reverse mortgage and does not belong to open-end line of credit.
★ Cases not for business or commercial purpose were mostly denied while some were approved. These might be for homes.

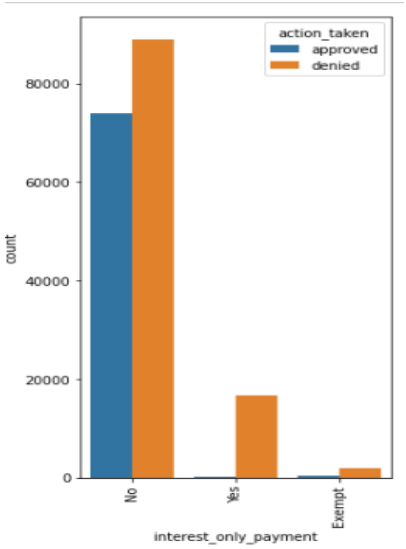
```
75-100%   66.17
50-75%    17.45
25-50%     8.73
0-25%      3.93
>100%      2.44
Exempt     1.27
Name: loan_to_value_ratio, dtype: float64
NUMBER OF CATEGORIES= 6
```



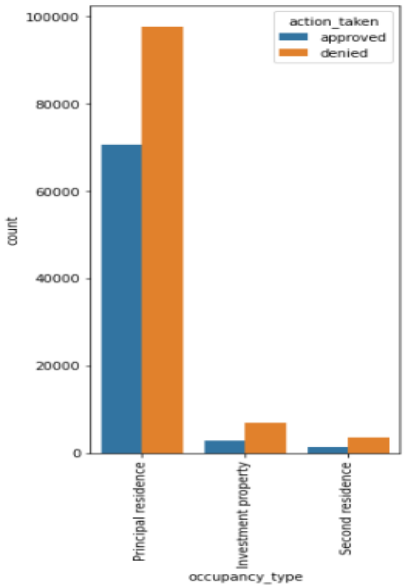
- ★ Only Loan amounts between 75%-100% of the property value have a good chance of getting loan approval whereas loan amounts with <75% or >100% of the property values are getting rejected.
- ★ Over 66% of customers applying have greater than 75% of loan to value ratio and 1% of customers are exempted from this very important standard.

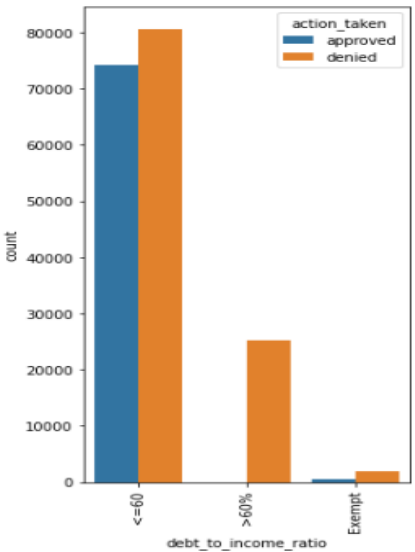
7. Property_value, interest_only_payment, occupancy type, debt_to_income_ratio:





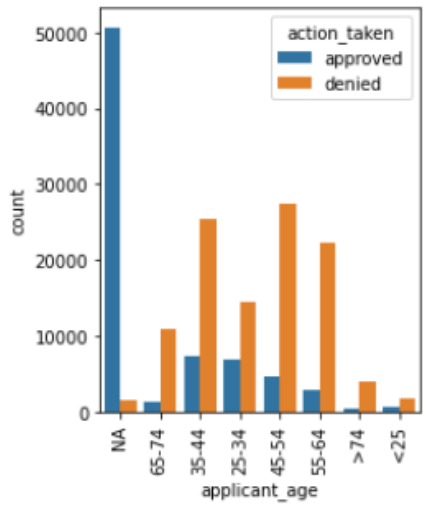
- ★ Properties with >1M value are also getting more rejections than approval ,which indicates that,'loan approval' is not dependent on the value of a property.
- ★ Exempted property value cases are mostly denied.
- ★ Higher % of people getting approved from their respective value brackets as the property value bracket keeps increasing.

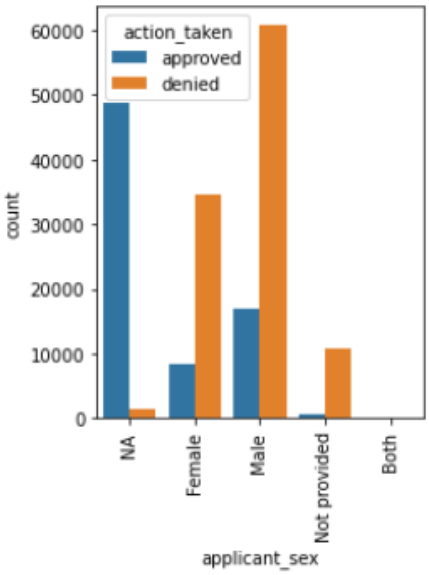




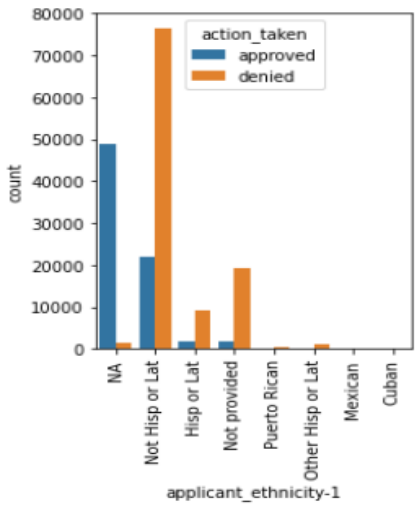
- ★ Loans are getting rejected,if the debt of the borrower is >60% of his/her income.
- ★ All of the credit scoring models are rejecting loan requests.Hence the borrowers are most likely to get their loan request rejected if their credit score is calculated using the credit scoring models.
- ★ Most of the people are taking a loan for principal residence of which a huge number are getting denied.

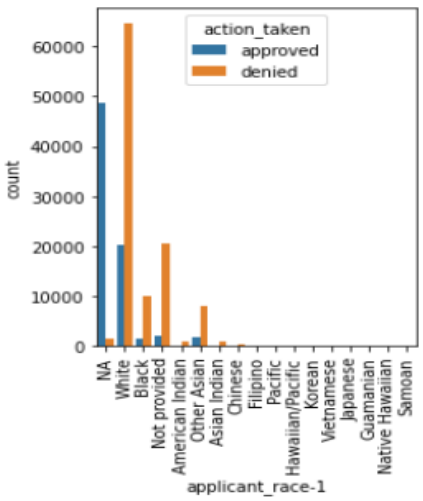
8. Applicant Age, sex, ethnicity and race:





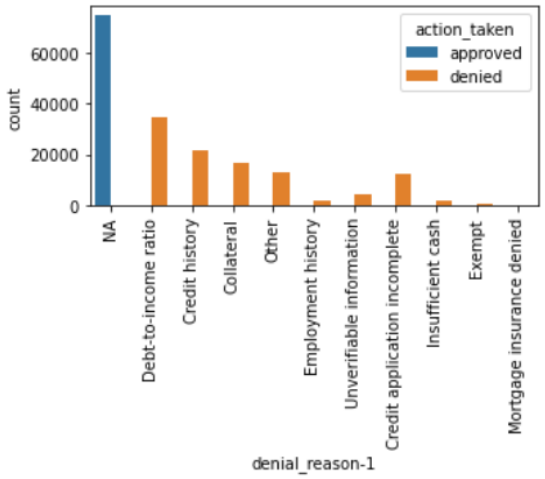
- ★ Applicant age does not seem a very very important factor because there are an huge amount of approved loans even where age is not applicable.
- ★ Of where it is applicable, the majority have been denied. This could perhaps be because of other parameters failing to meet the conditions.
- ★ Applicant sex is not an important factor because a lot of loans have been approved irrespective of gender. Where they have asked for it and it has not been provided, the loans have been denied. Negligible cases of non-binary genders.

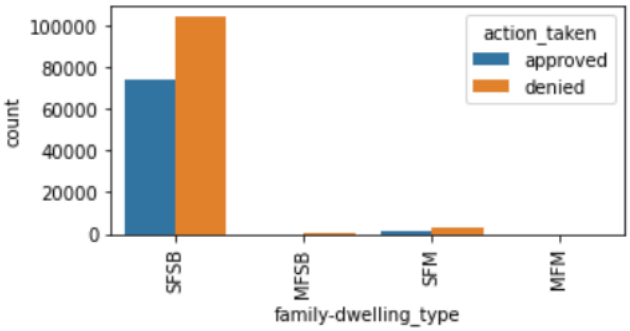




- ★ Applicant ethnicity is not a big parameter although significant because even without the ethnicity not being applicable, lots of loans have been approved.
- ★ Loans have been denied majorly to almost all the races but if they are especially hispanic or latin, no loans have been approved.
- ★ Loans of Indians have been denied and none approved and the number of denials is also high for the other specified races.
- ★ The number of loan approvals are high where race has not been applicable.

9. Reason for denial, family dwelling type:

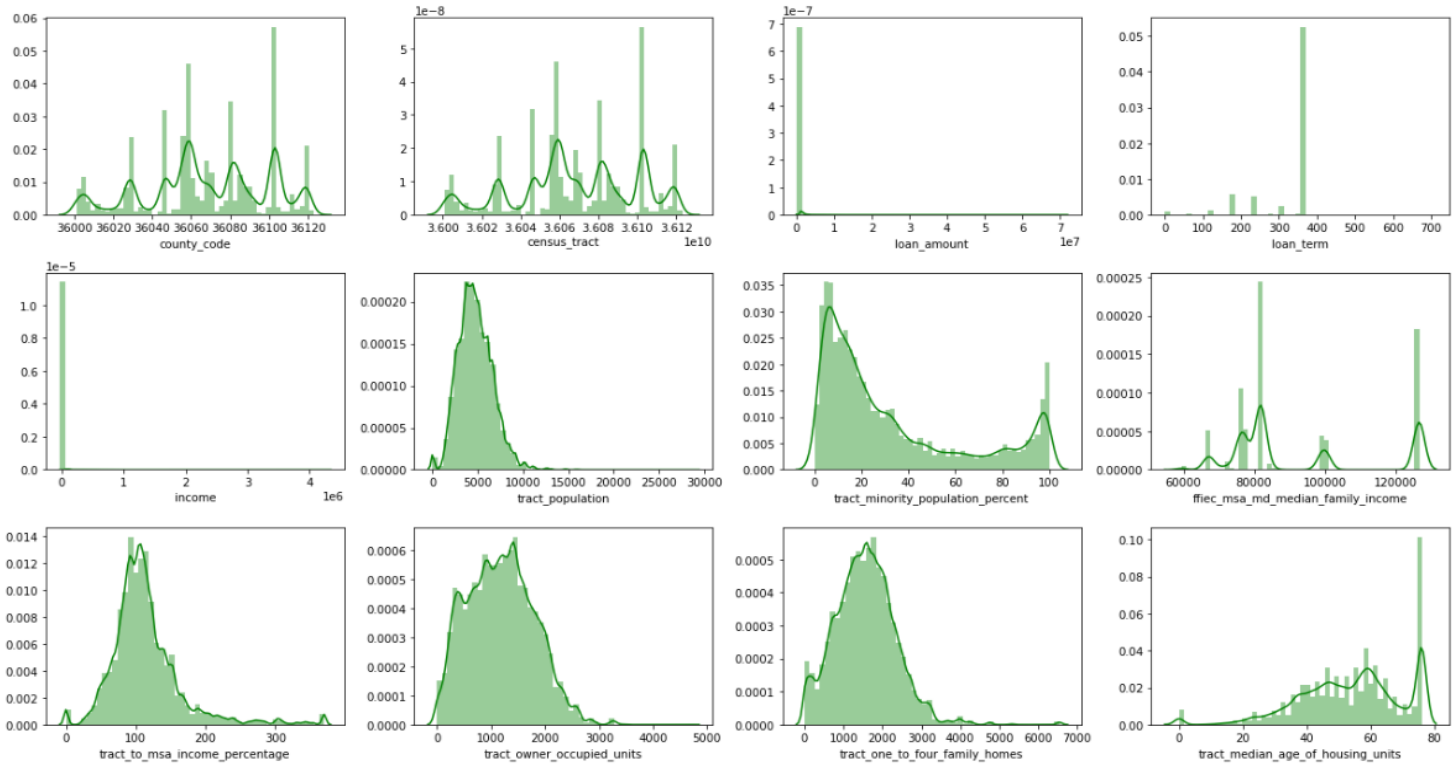




- ★ The highest number of denials come from the debt-to-income ratio meaning it is an important parameter.
- ★ Most people are in an SFSB dwelling type and none in MFM.
- ★ All those in MFSB have been denied loans.

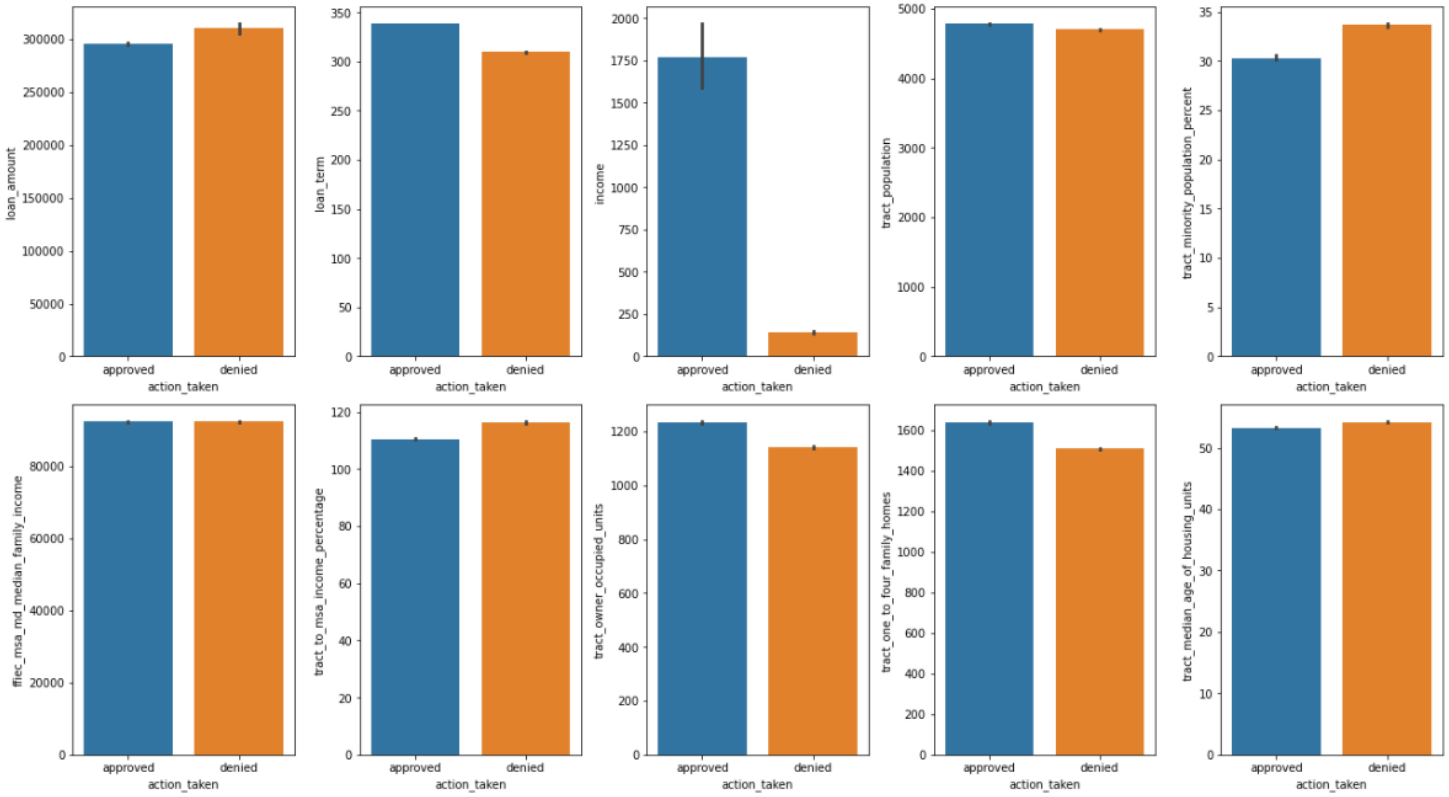
B. Numerical:

1. Check the Distribution:



- ★ loan_amount and income columns are very heavily skewed.
- ★ tract_minority_population_percent,tract_to_msa_income_percentage are highly positive skewed.
- ★ loan_term is higly negative skewed.
- ★ tract_median_age_of_housing_units,tract_one_to_four_family_homes,tract_population,ffiec_msa_md_median_family_income are slightly skewed.
- ★ tract_owner_occupied_units is the only column with near normal distribution.

2. Numeric vs Target:

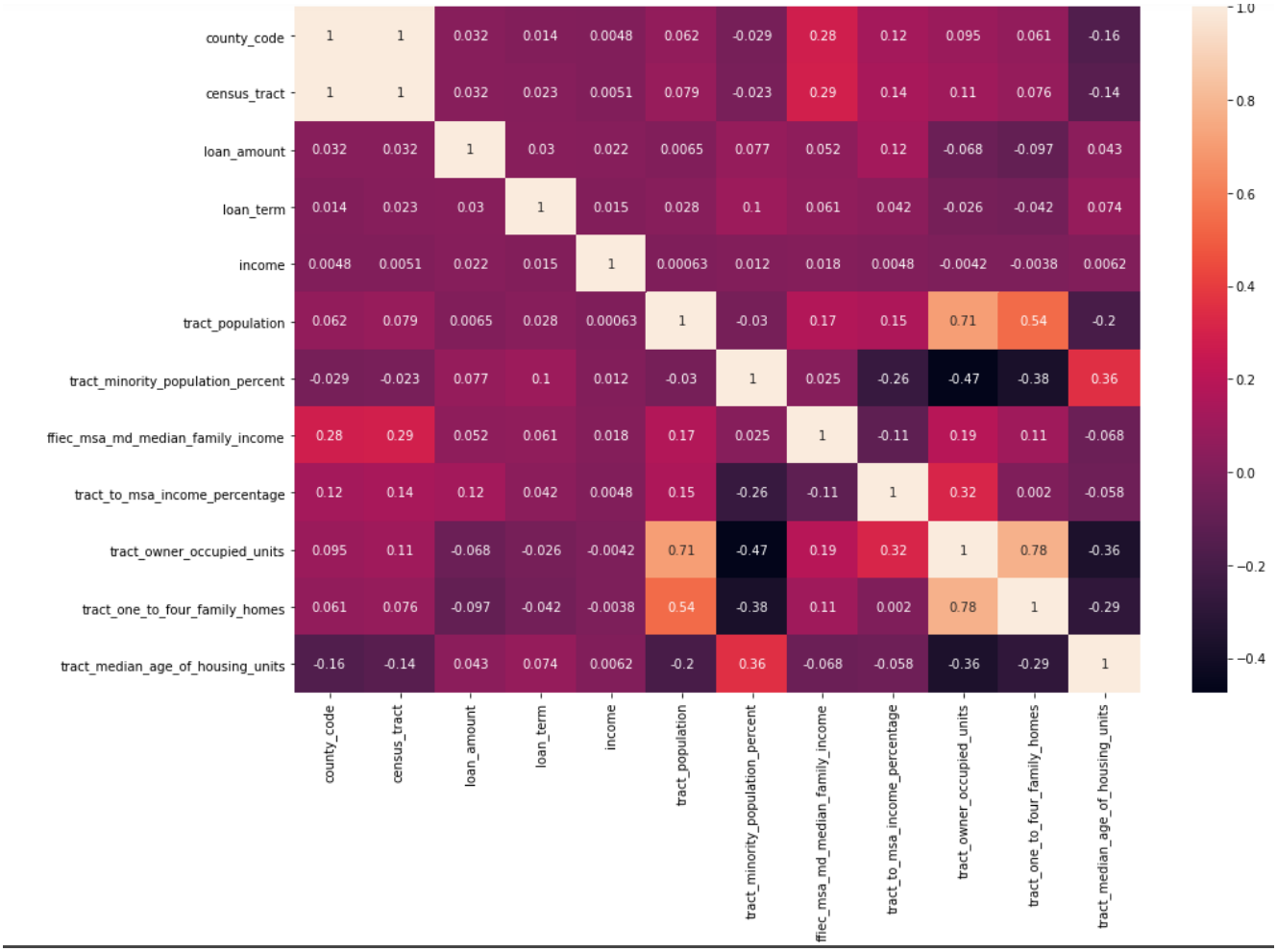


Findings: *The following findings are infered based on average values.*

- ★ Maximum variation comes from the Income and there is almost no judgement that we can form from most of the other numeric features.
- ★ Loan term and loan amount can also be of help in separating the target.
- ★ This, however, does not mean we can eliminate features just yet.
- ★ The mean loan_amount for denied loans are higher,meaning borrowers whose loans are denied were requesting higher loan_amounts than those for whom the loan is approved.

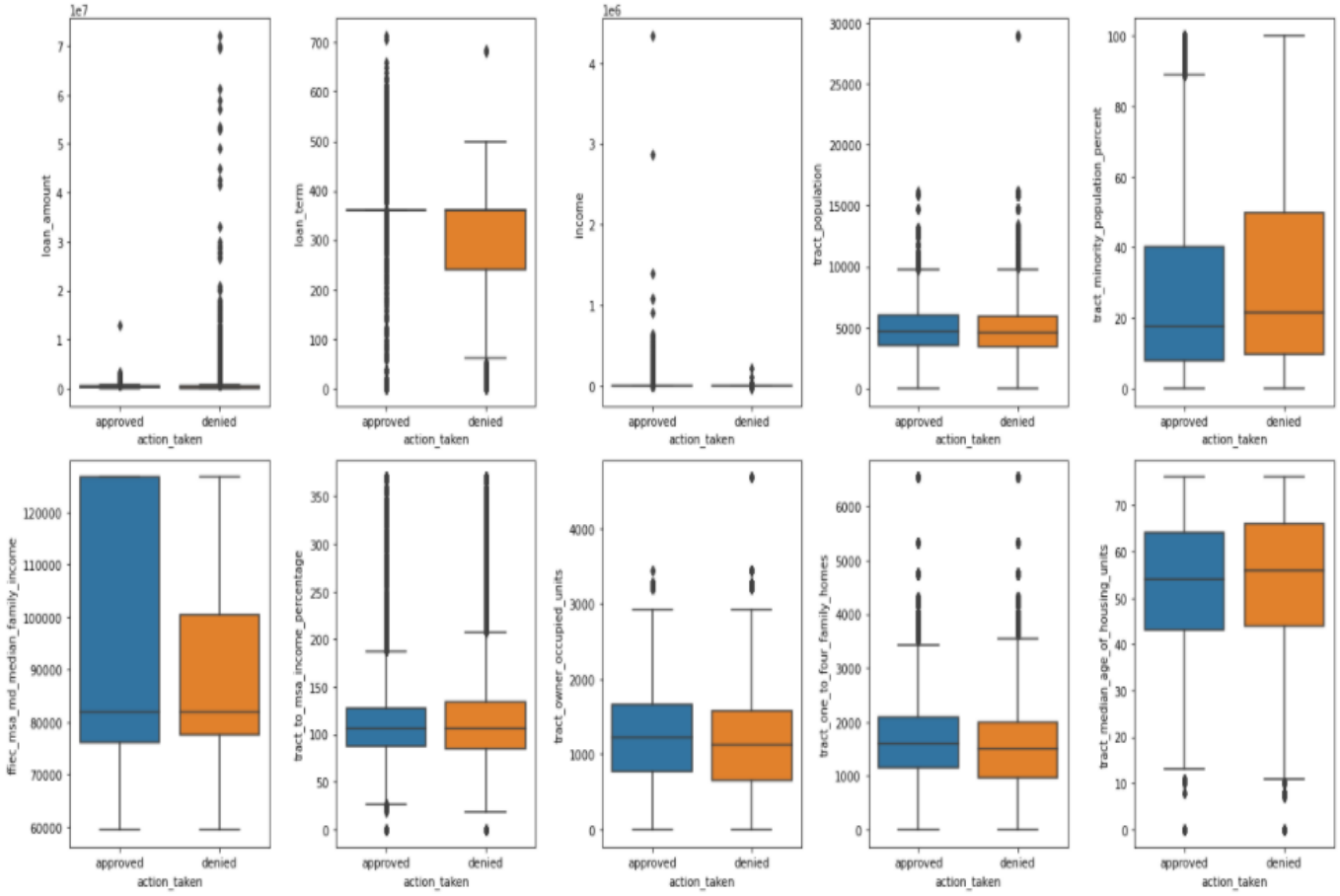
- ★ Income of a borrower has great influence on determining whether a loan is approved or not.The mean income for borrowers whose application is denied is very low.
- ★ The average percentage for minority population is high for denied loans,implying that there is some level of discriminization in tracts where the percentage of minority population is high.

3. Multicollinearity:



- ★ Country code and census tract have a correlation of 1 which means they represent one another.
- ★ tract one to four family homes has multicollinearity with tract population and tract owner occupied units.
- ★ tract population also has a multicollinearity with tract owner occupied units.
- ★ Conclusion is that tract one to four family homes, tract population and tract owner occupied units are features that are not independent from one another. We may consider developing a new column out of the 3 or consider dropping one of them.

4. Outlier Detection:



- ★ except ffiec_msa_md_median_family_income, all the numeric features have a good amount of outliers. It is to be decided whether such data must be capped or eliminated.
- ★ As a best practice, to avoid data leakage, we can eliminate the top and bottom 1% of the data (trim) and then cap the rest of the outliers.
- ★ Loan term, ffiec_msa_md_median_family_income are features that can be considered as some of the best indicators of action taken.

5. Statisticalsignificanceofvariables:

```
Hypothesis:
H0: county_code does not have any impact on action taken
H1: county_code has significant impact on action taken

Pvalue= 1.7245450238736588e-36
Reject null hypothesis, county_code has an impact on target
--------------------------------------------------------------------------------

Hypothesis:
H0: census_tract does not have any impact on action taken
H1: census_tract has significant impact on action taken

Pvalue= 1.8052653872841565e-28
Reject null hypothesis, census_tract has an impact on target
--------------------------------------------------------------------------------

Hypothesis:
H0: loan_amount does not have any impact on action taken
H1: loan_amount has significant impact on action taken

Pvalue= 3.944192074405708e-06
Reject null hypothesis, loan_amount has an impact on target
--------------------------------------------------------------------------------

Hypothesis:
H0: loan_term does not have any impact on action taken
H1: loan_term has significant impact on action taken

Pvalue= 0.0
Reject null hypothesis, loan_term has an impact on target
--------------------------------------------------------------------------------

Hypothesis:
H0: income does not have any impact on action taken
H1: income has significant impact on action taken

Pvalue= 2.2292448306375457e-91
Reject null hypothesis, income has an impact on target
Hypothesis:
H0: tract_population does not have any impact on action taken
H1: tract_population has significant impact on action taken

Pvalue= 5.243440402175278e-19
Reject null hypothesis, tract_population has an impact on target
--------------------------------------------------------------------------------

Hypothesis:
H0: tract_minority_population_percent does not have any impact on action taken
H1: tract_minority_population_percent has significant impact on action taken

Pvalue= 4.491909381739524e-118
Reject null hypothesis, tract_minority_population_percent has an impact on target
--------------------------------------------------------------------------------

Hypothesis:
H0: ffiec_msa_md_median_family_income does not have any impact on action taken
H1: ffiec_msa_md_median_family_income has significant impact on action taken

Pvalue= 0.8849762378529389
Fail to reject null hypothesis, ffiec_msa_md_median_family_income has no impact on target
--------------------------------------------------------------------------------

Hypothesis:
H0: tract_to_msa_income_percentage does not have any impact on action taken
H1: tract_to_msa_income_percentage has significant impact on action taken

Pvalue= 4.8609754661231156e-127
Reject null hypothesis, tract_to_msa_income_percentage has an impact on target
--------------------------------------------------------------------------------

Hypothesis:
H0: tract_owner_occupied_units does not have any impact on action taken
H1: tract_owner_occupied_units has significant impact on action taken

Pvalue= 7.165992236795271e-221
Reject null hypothesis, tract_owner_occupied_units has an impact on target
```

- ★ We ran an anova test to test the significance of all numeric features to the target feature which happens to be categorical.
- ★ From the test results we could say that all the variables have the p-value less than 0.05, which means they reject the null hypothesis which indirectly tells us that there is strong influence of them on atcion_taken i.e., target variable.
- ★ ffiec_msa_md_median_family_income is the only numeric column to not be statistically significant for action taken. It can be dropped for the purpose of building the model.

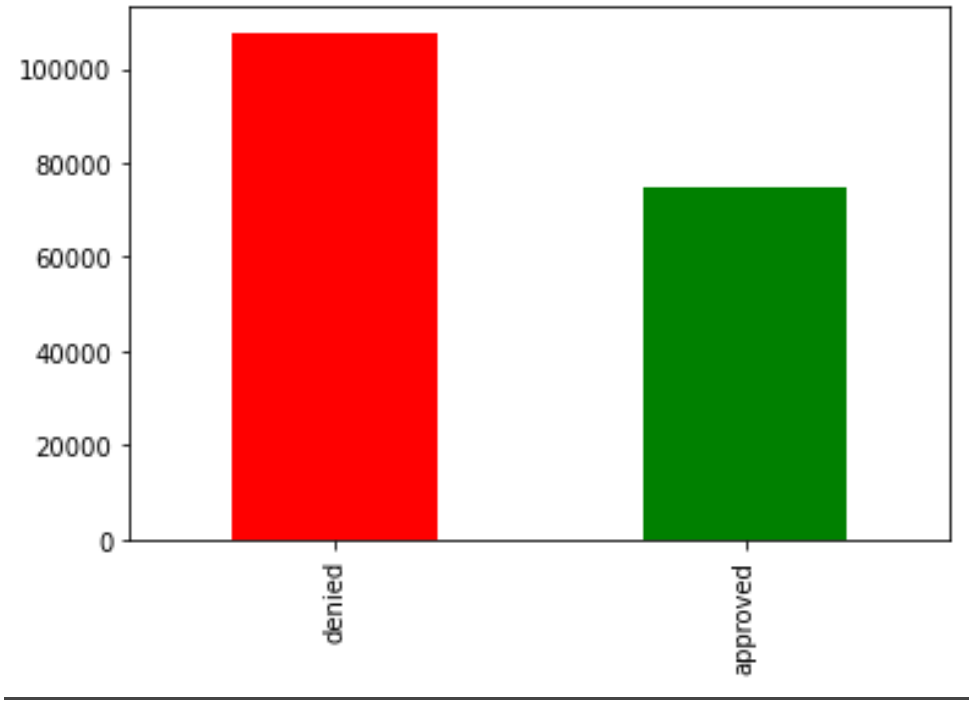
- Chi-square test for categorical variables:

```
Reject null hypothesis, lei has an impact on target
Reject null hypothesis, derived_msa-md has an impact on target
Reject null hypothesis, conforming_loan_limit has an impact on target
Reject null hypothesis, action_taken has an impact on target
Reject null hypothesis, purchaser_type has an impact on target
Reject null hypothesis, loan_type has an impact on target
Reject null hypothesis, loan_purpose has an impact on target
Reject null hypothesis, lien_status has an impact on target
Reject null hypothesis, reverse_mortgage has an impact on target
Reject null hypothesis, open-end_line_of_credit has an impact on target
Reject null hypothesis, business_or_commercial_purpose has an impact on target
Reject null hypothesis, loan_to_value_ratio has an impact on target
Reject null hypothesis, negative_amortization has an impact on target
Reject null hypothesis, interest_only_payment has an impact on target
Reject null hypothesis, balloon_payment has an impact on target
Reject null hypothesis, other_nonamortizing_features has an impact on target
Reject null hypothesis, property_value has an impact on target
Reject null hypothesis, occupancy_type has an impact on target
Reject null hypothesis, total_units has an impact on target
Reject null hypothesis, debt_to_income_ratio has an impact on target
Reject null hypothesis, applicant_credit_score_type has an impact on target
Reject null hypothesis, co-applicant_credit_score_type has an impact on target
Reject null hypothesis, applicant_ethnicity-1 has an impact on target
Reject null hypothesis, co-applicant_ethnicity-1 has an impact on target
Reject null hypothesis, applicant_race-1 has an impact on target
Reject null hypothesis, co-applicant_race-1 has an impact on target
Reject null hypothesis, applicant_sex has an impact on target
Reject null hypothesis, co-applicant_sex has an impact on target
Reject null hypothesis, applicant_age has an impact on target
Reject null hypothesis, co-applicant_age has an impact on target
Reject null hypothesis, submission_of_application has an impact on target
Reject null hypothesis, initially_payable_to_institution has an impact on target
Reject null hypothesis, aus-1 has an impact on target
Reject null hypothesis, denial_reason-1 has an impact on target
Reject null hypothesis, family-dwelling_type has an impact on target
Reject null hypothesis, high_cost_mortgage has an impact on target
```

- ★ We are performing the chi-square test between the independent categorical variable with the dependent variable i.e. action_taken.

- ★ From the above screenshot, it is evident for all the variables the p-value is less than 0.05. Hence, we can conclude that all the variables are significant i.e., there is a relationship between those variables
- ★ This proves all our explorations in the past where we have found varied results with respect to the target variable.

6. **Check Class Imbalance:**



- ★ 'Approved' is the minority class.
- ★ The difference is not alarming and there is no need for treatment of class imbalance.

APPROACH AND STEPS TAKEN TO SOLVE THE BUSINESS PROBLEM:

Step 1: Business Understanding

- Our customer here was the prospective applicant of a home loan.
- There is already a systematic means of dataset available that comes from a reliable government website.
- This was to impact the banking and financial services industry as it involves the process of home loans and BFSI companies that provide these services.
- The project was under the purview of the HMDA Act. It would assist the purpose of the HMDA Act.

Step 2: Data Understanding

- We collected the demographics of the applicant including age, ethnicity and race from the dataset as well as the reason for application, financial information of the applicant and co-applicant.
- Any other home mortgage information to assist the data understanding
- The income of the applicant, income of the population, income geography wise (mean and median) was analysed.
- The effect of categorical variables with respect to the target is to be analysed.
- There is presence of ‘Not available’ and ‘Not applicable’ both. The former is a disguised null value and the latter is an important indicator of whether there is bias in lending or not.

Step 3: Data Preparation

- Filter out the features with over 50% of null values present.
- Filter out features with no variability
- Identify the target feature as action_taken : classes are ‘approved’ and ‘denied’
- Filter out features with no statistical significance to the target.
- Check the presence of false 0 values and treat them as null.
- Impute the null values.
- Check for outliers and treat them
- Check for multicollinearity, normality, class imbalance and treat if defects are present.
- Decide how to encode categorical features and do the same.
- Scale the data and split it in a train and test of size 30%.

STEP 4: Modeling

- A classification model would be best suited to the business objective.
- Depending upon the outliers, the nature of the dataset after exploration, the number of features, the classification algorithm was to be selected.
- For every model built, the model metrics were recorded for analysis and comparison.
- The logistic regression full model was the base model.
- Decision tree Algorithms, Random Forest and other ensemble algorithms were tried. After various loops of trials, the random forest classifier model was observed to be the winner model.
- Since the dataset consists of nearly 50 features,Recursive Feature Selection(RFE) technique was used to perform feature selection.

STEP 5: Evaluation

- Evaluate the accuracy score and error metrics.
- Plot the roc_auc curve and confusion matrix and evaluate the false positives.
- Analyse the classification report of both train and test set.
- Rank the metrics of the models.
- Re-evaluate the process on which the models are built.

MODEL BUILDING:

After performing all the data preprocessing ,the shape of the final dataset is 182272 rows and 47 columns with 9 numerical columns and 36 categorical columns. The categorical columns must be converted to numeric for building the model,for that purpose feature encoding techniques are used.

Removing some more columns:

```
In [83]: #since these are details that a customer can know only after knowing his/her loan status,these columns should be removed
         x=df1.drop(['lei','aus-1','co-applicant_credit_score_type','applicant_credit_score_type','action_taken','denial_reason-1'],axis=1
         y=df1['action_taken']
```

Feature Encoding:

Since the dataset consist of 36 categorical columns one-hot encoding will significantly increase the number of columns, since the number of unique values for each of these columns are less, Frequency Encoding is used to encode the categorical variables.

```
In [77]: #frequency encoding categorical columns
         for i in s:
             tab=df1[i].value_counts(normalize=True)*100
             df1[i]=df1[i].map(tab)
         df1
```

Where ‘’s’’ is a list of categorical features.

Feature Scaling:

Since the numeric and encoded features have different distributions, these features should be normalized,hence Standard Scaler is used for scaling all the features.

]:

	derived_msa-md	census_tract	conforming_loan_limit	purchaser_type	loan_type	loan_purpose	lien_status	reverse_mortgage	open-end_line_of_credit	bu
0	0.45	0.58	-5.22	0.60	-2.27	0.54	0.44	0.13	0.50	
1	0.45	0.05	0.19	-1.65	-2.27	1.07	0.44	0.13	0.50	
2	0.45	0.80	0.19	-1.65	-1.69	0.54	0.44	0.13	0.50	
3	-1.00	0.12	0.19	-1.65	-2.27	1.07	0.44	0.13	0.50	
4	-1.00	0.21	0.19	-1.65	-2.27	1.07	0.44	0.13	0.50	

Base Model:

The base model was built on Logistic regression(Logit) and the following result was obtained:

Logit Regression Results

Dep. Variable:	action_taken	**No. Observations:**	127590
Model:	Logit	**Df Residuals:**	127548
Method:	MLE	**Df Model:**	41
Date:	Wed, 08 Sep 2021	**Pseudo R-squ.:**	0.9788
Time:	17:48:00	**Log-Likelihood:**	-1830.7
converged:	False	**LL-Null:**	-86337.
Covariance Type:	nonrobust	**LLR p-value:**	0.000

Possibly complete quasi-separation: A fraction 0.81 of observations can be perfectly predicted. This might indicate that there is complete quasi-separation. In this case some parameters will not be identified.

Other Models:

Similarly various models were built on other Machine Learning algorithms like Decision Tree,Random Forest,GuassianNB,Gradient Boosting and Xgboost.The following train and test scores were obtained:

```
Logreg train: 0.9972803511246963
Logreg test: 0.9974946051717202
DT train: 1.0
DT test: 0.9998536995720713
RF train: 1.0
RF test: 0.9998719871255624
GaussNB train: 0.9924680617603261
GaussNB test: 0.9931421674408397
GradBoost train: 0.9998667607179246
GradBoost test: 0.9998536995720713
[17:49:40] WARNING: C:/Users/Administrator/workspace
0, the default evaluation metric used with the objec
t eval_metric if you'd like to restore the old behav
XGB train: 1.0
XGB test: 0.9999634248930178
```

Both the train and test accuracy for all the models are above 0.99 which means that all the models were able to predict the status of a loan with 99% accuracy,which is remarkable.

Feature Selection:

Since the dataset consist of 41 features and the models are achieving good accuracy,reducing these features can be useful in building a deployable model(although we are not deploying the model).
For this purpose, Recursive Feature Selection is used to perform feature selection.

```
rfe_feat
```

```
Index(['purchaser_type', 'loan_type', 'loan_purpose',
       'open-end_line_of_credit', 'business_or_commercial_purpose',
       'loan_to_value_ratio', 'income', 'debt_to_income_ratio',
       'co-applicant_ethnicity-1', 'co-applicant_sex', 'applicant_age',
       'co-applicant_age', 'submission_of_application',
       'initially_payable_to_institution', 'high_cost_mortgage'],
      dtype='object')
```

The above features are obtained by performing Recursive feature selection and models were built with these features and the following results were obtained.

```
Logreg train: 0.9942863860804139
Logreg test: 0.9944405837387075
DT train: 0.9999294615565483
DT test: 0.9997073991441425
RF train: 0.9999294615565483
RF test: 0.9998354120185802
GaussNB train: 0.9932518222431225
GaussNB test: 0.9936542189385904
GradBoost train: 0.999796222274473
GradBoost test: 0.9998719871255624
[18:38:17] WARNING: C:/Users/Adminis
0, the default evaluation metric use
t eval_metric if you'd like to restc
XGB train: 0.9998824359275805
XGB test: 0.9998354120185802
```

Cross validating models:

Cross validation is necessary for validating the models with different sets of data for a number iterations,to evaluate how well the model is performing on different sets of data.The model with least bias and variance error is supposed to be the best model.

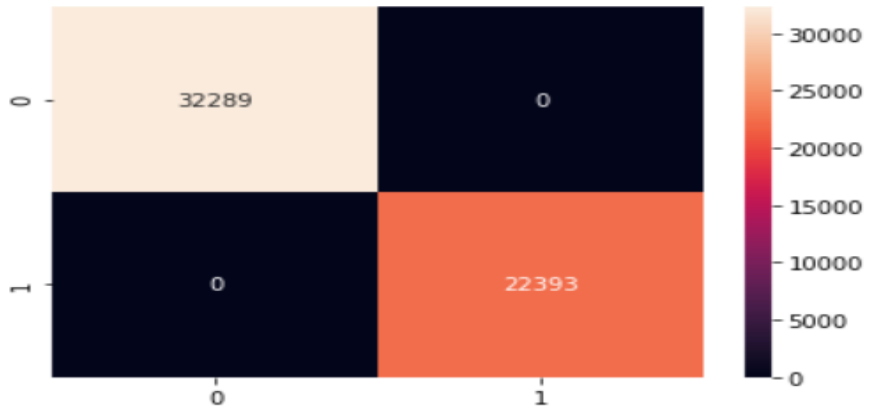
```
Logreg - bias error : 0.005666588290618391
Logreg - variance error : 0.00015486204217321573
DT - bias error : 0.00021945293518288445
DT - variance error : 4.703595114532746e-05
RF - bias error : 0.00019594012069923394
RF - variance error : 2.4789539922145715e-05
GaussNB - bias error : 0.006693314523081639
GaussNB - variance error : 0.00031699453078209256
GradBoost - bias error : 0.00016458970138732987
GradBoost - variance error : 4.570821982492264e-05
[18:40:28] WARNING: C:/Users/Administrator/workspace/xgboos
0, the default evaluation metric used with the objective 'b
t eval_metric if you'd like to restore the old behavior.
[18:40:30] WARNING: C:/Users/Administrator/workspace/xgboos
0, the default evaluation metric used with the objective 'b
t eval_metric if you'd like to restore the old behavior.
[18:40:31] WARNING: C:/Users/Administrator/workspace/xgboos
0, the default evaluation metric used with the objective 'b
t eval_metric if you'd like to restore the old behavior.
[18:40:33] WARNING: C:/Users/Administrator/workspace/xgboos
0, the default evaluation metric used with the objective 'b
t eval_metric if you'd like to restore the old behavior.
[18:40:34] WARNING: C:/Users/Administrator/workspace/xgboos
0, the default evaluation metric used with the objective 'b
t eval_metric if you'd like to restore the old behavior.
XGB - bias error : 0.00020377772552715445
XGB - variance error : 4.57100114048286e-05
```

Random Forest model has achieved the least bais and variance error.

MODEL EVALUATION:

- After running several classification algorithms,Random forest seemed to perform slightly better than the rest of the models. Hence Random Forest was used as the final model.
- The model accuracy didn't reduce after performing feature selection and so a Random Forest model was built upon the reduced features.
- Various evaluation metrics were also used to evaluate the performance of the model.

Confusion Matrix:



The above heatmap shows that the model has perfectly classified both the class approved(1) and denied(0).

Classification Report:

```
[191]: print(classification_report(y_test,fin_pred))
```

```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     32289
           1       1.00      1.00      1.00     22393

    accuracy                           1.00     54682
   macro avg       1.00      1.00      1.00     54682
weighted avg       1.00      1.00      1.00     54682
```

Looking at the 'f1-score' of weighted average,the model has achieved 100% accuracy in determining the loan status of the customers.

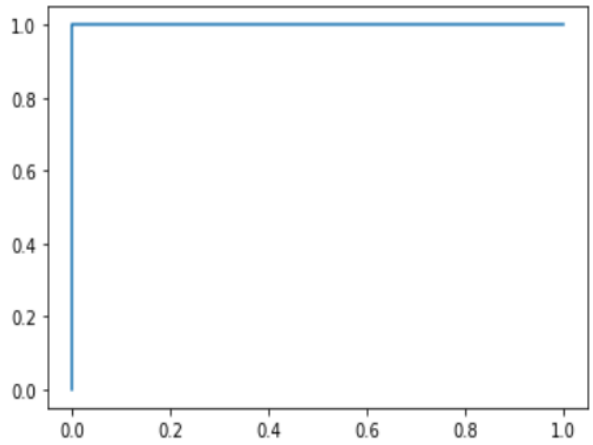
Roc_auc score:

The roc auc score represents the area under the curve,the closer the roc_auc_score to 1.0, the better is the model performance.

```
roc_auc_score(y_test,fin_prob[:,1])
```

```
[194… 1.0
```

```
[193]: plt.plot(fpr,tpr)
       plt.show()
```



From the above plot,it is evident that the model has a perfect area under the curve.
From the above evaluations,it is evident that the model is providing accurate predictions and have aced all the evaluation metrics.

Implications:

- Our work suggests that applying a machine learning approach to a particular Home Mortgage Disclosure Act (HMDA) dataset, reduces the stress of going through all the details of a particular
- individual to check whether a person can receive a loan based on his financial,personal and geographic factors.Thereby,this model can be used as preliminary procedure by a financial
- institution instead of hiring a huge team. This is useful for any newly established for profit financial institution to use this particular model to automate the whole preliminary procedure ,because it is not able to invest huge amounts in the data analysis of each and every individual's data.
- By using this model,they can target only those individuals who got the approved status after their information has gone through this model,thereby saving huge time also and investing their activities in the particular zone in their data where the status has got approved.
- As the model is using only those features which have gone through statistical analysis ,each feature is 95 percent confident with regard to the target.

Limitations:

- The model was built on dataset that has customer information from the banks that belong to the state of New York. Hence,there is no guarantee that the model will make better predictions for banks that belong to other states, since each state has its own demographics and financial standards.
- Additionally, from the model perspective, this has been built considering many parameters hence the risk of an unrealistic accuracy remains. There may be presence of overfitting that can also go unnoticed. The volume of the data as well can lead to such happening.
- Many parameters may have been statistically significant to the target but it is possible that some of these actually have no relevance in giving out a loan and can be pure coincidence.

Conclusion:

- The main idea of the project was to brainstorm and build a tool that would solve the entire purpose of the HMDA Act.
- By working on this project, we have successfully built something of meaning and as students that makes us happy because we are able to use whatever we have learnt and make something for a good cause.
- We have learnt to work as a strong team on this and that has improved our communication skills, presentation skills, report writing skills and much more.
- We got a chance to work on real life data that was huge in size and to develop insights on that, opened up our analytical minds a lot more. We understood how to efficiently read data as well as understand what the unclean data also speaks. We were then able to clean it and turn data into information.