Data Mining Project (CMPE 255)

HDMA WASHINGTON STATE HOME LOANS, 2016

December 2021

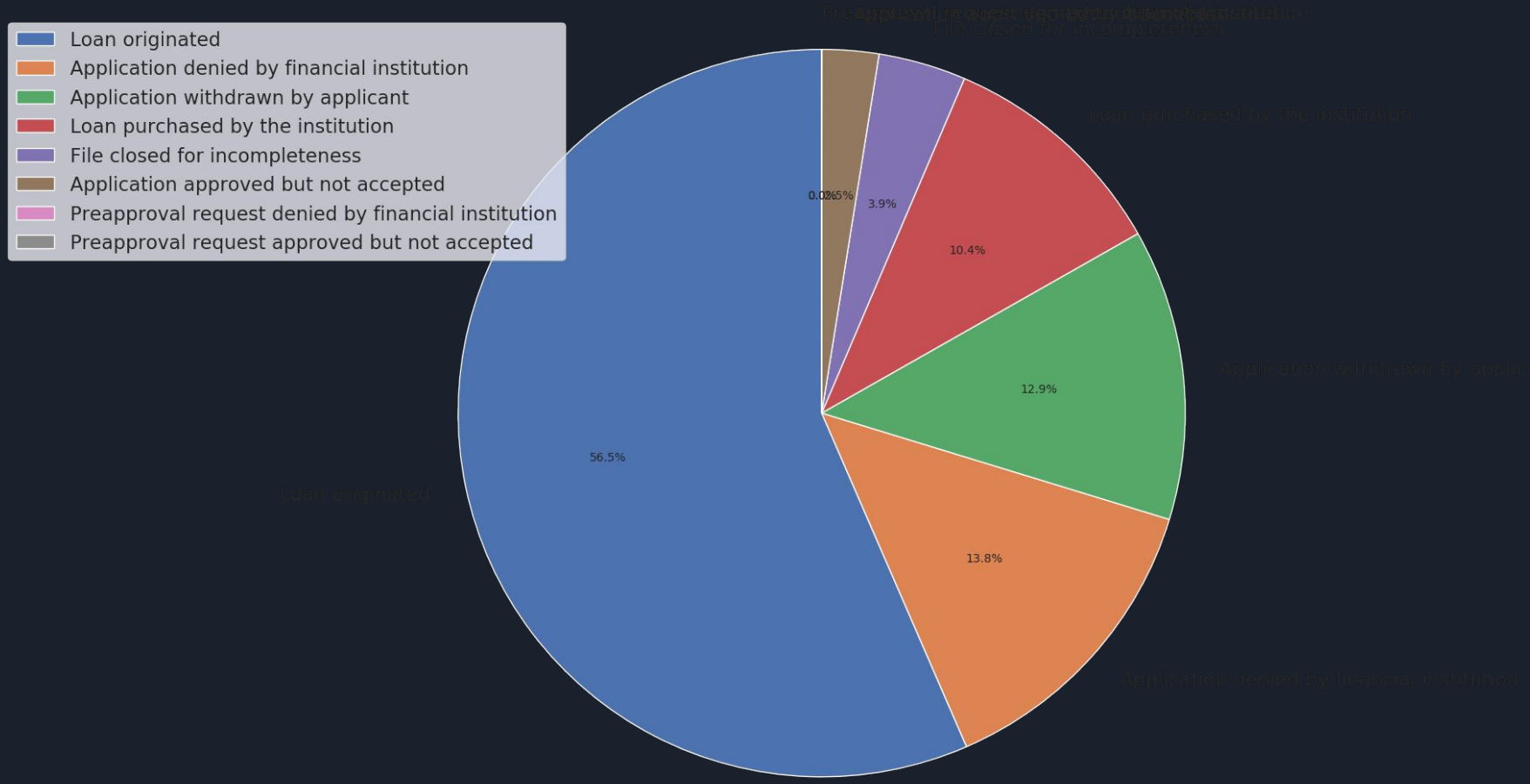
Ishan Hitesh Malkan (015281968)
Romil Viral Shah (015250599)
Pallav Parikh (015275481)

Introduction

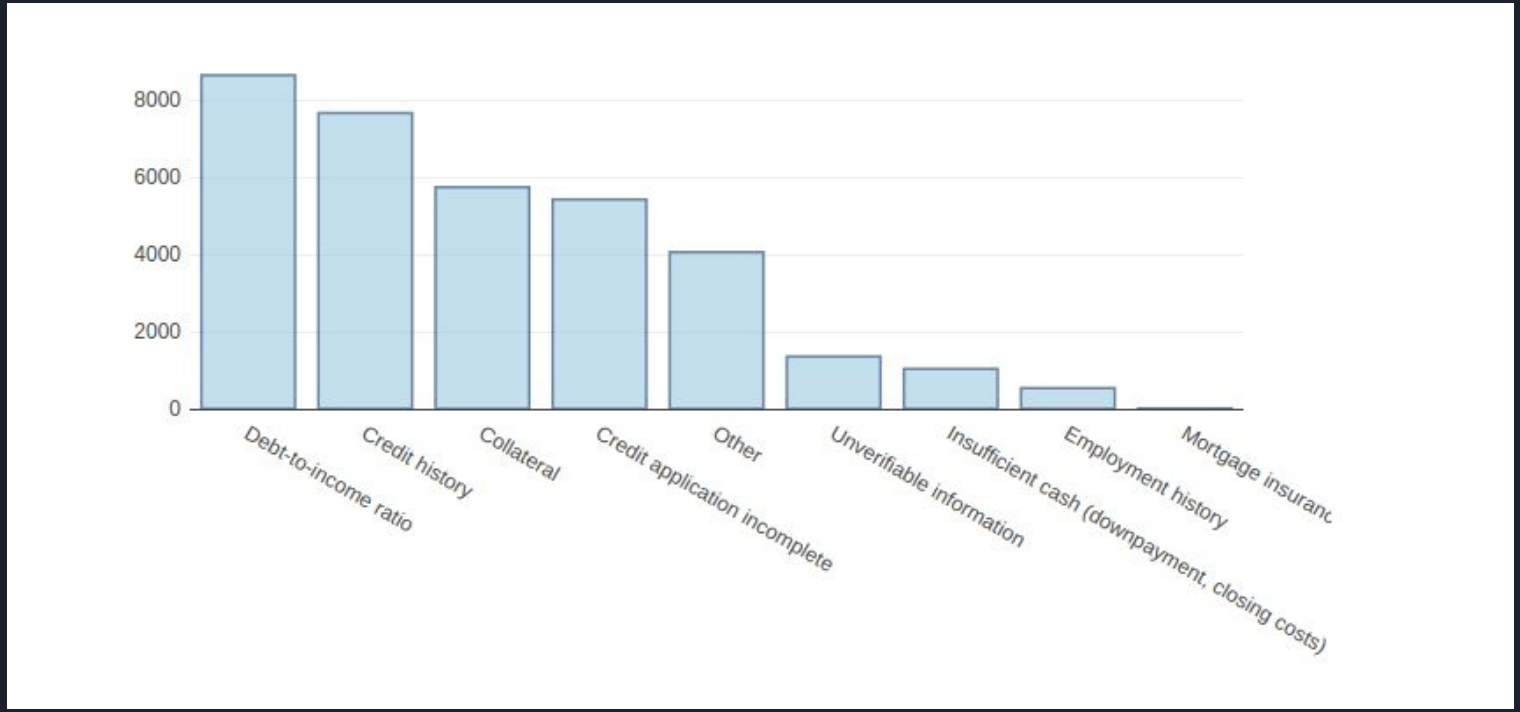
- Explore the data for Washington State Home Loans, 2016.
- Performing preprocessing, exploratory data analysis and classification using multiple algorithms.
- Implementation of a predictive model that loan will be approved or not for the person

Dataset Description

```
Index(['tract_to_msamd_income', 'rate_spread', 'population',
       'minority_population', 'number_of_owner_occupied_units',
       'number_of_1_to_4_family_units', 'loan_amount_000s',
       'hud_median_family_income', 'applicant_income_000s', 'state_name',
       'state_abbr', 'sequence_number', 'respondent_id', 'purchaser_type_name
       'property_type_name', 'preapproval_name', 'owner_occupancy_name',
       'msamd_name', 'loan_type_name', 'loan_purpose_name', 'lien_status_name
       'hoepa_status_name', 'edit_status_name', 'denial_reason_name_3',
       'denial_reason_name_2', 'denial_reason_name_1', 'county_name',
       'co_applicant_sex_name', 'co_applicant_race_name_5',
       'co_applicant_race_name_4', 'co_applicant_race_name_3',
       'co_applicant_race_name_2', 'co_applicant_race_name_1',
       'co_applicant_ethnicity_name', 'census_tract_number', 'as_of_year',
       'application_date_indicator', 'applicant_sex_name',
       'applicant_race_name_5', 'applicant_race_name_4',
       'applicant_race_name_3', 'applicant_race_name_2',
       'applicant_race_name_1', 'applicant_ethnicity_name', 'agency_name',
       'agency_abbr', 'action_taken_name'],
```


Loan originated
Application denied by financial institution
Application withdrawn by applicant
Loan purchased by the institution
File closed for incompleteness
Application approved but not accepted
Preapproval request denied by financial institution
Preapproval request approved but not accepted
56.5%
13.8%
12.9%
10.4%
3.9%
0.0%
2.5%

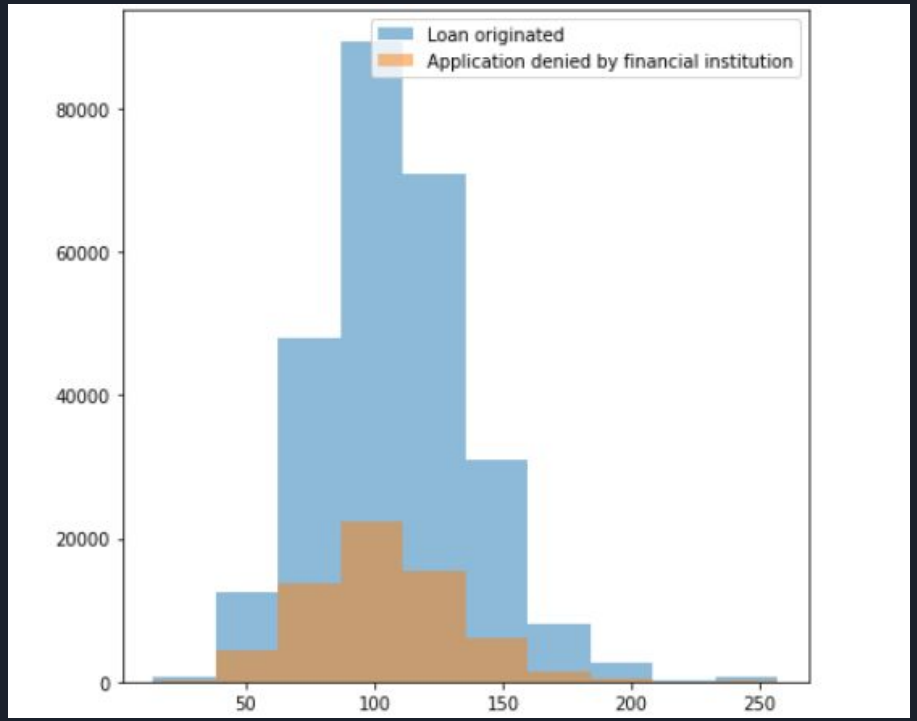
Reasons For Loan Denial



Preprocessing on data

- Removing the extra columns such as names / constant columns etc. which won't be required in further analysis and modelling.
- Removing columns with greater than 95% nulls (not significant in further analysis)
- Separating the columns with numerical values and the ones with categorical values.
- Imputing the missing values using various strategies
- Binning the numerical features
- One hot encoding the features

Analysing the numerical features



- Analysing distribution of feature grouping by action taken

	tract_to_msamd_income bin lower end	tract_to_msamd_income bin upper end	tract_to_msamd_income Loan originated Freq	tract_to_msamd_income Application denied by financial institution Freq	tract_to_msamd_income Loan originated % Freq	tract_to_msamd_income Application denied by financial institution % Freq
0	14.050000	38.359002	791.0	176.0	0.300019	0.274974
1	38.359002	62.668003	12412.0	4299.0	4.707756	6.716558
2	62.668003	86.977005	47934.0	13619.0	18.180922	21.277693
3	86.977005	111.286006	89230.0	22384.0	33.844112	34.971721
4	111.286006	135.595007	70765.0	15382.0	26.840508	24.032122
5	135.595007	159.904009	30967.0	6105.0	11.745496	9.538168
6	159.904009	184.213010	8043.0	1417.0	3.050635	2.213855
7	184.213010	208.522012	2746.0	491.0	1.041532	0.767116
8	208.522012	232.831013	191.0	30.0	0.072445	0.046871
9	232.831013	257.140015	571.0	103.0	0.216575	0.160922

- Checking the numerical data for loan approval.

Checking the percentage of NULL values

	0
tract_to_msamd_income	0.131814
population	0.130742
minority_population	0.130742
number_of_owner_occupied_units	0.133314
number_of_1_to_4_family_units	0.130957
loan_amount_000s	0.000000
hud_median_family_income	0.129885
applicant_income_000s	13.295654
census_tract_number	0.129885
application_date_indicator	0.000000
action_taken_name	0.000000

Handling nulls in numerical values

- If the feature has <5% nulls, impute it with median
- If the feature has >=5% nulls, make null a separate category.

```
cols with less than 5% nulls:  ['tract_to_msamd_income', 'population', 'minority_population', 'number_of_owner_occupied_units', 'number_of_1_to_4_family_units', 'loan_amount_000s', 'hud_median_family_income', 'census_tract_number', 'application_date_indicator', 'action_taken_name']
cols with more than 5% nulls:  ['applicant_income_000s']
```

Null Imputed Data Max Values

tract_to_msamd_income	257.140015
population	13025.0
minority_population	94.790001
number_of_owner_occupied_units	2997.0
number_of_1_to_4_family_units	5893.0
loan_amount_000s	99999
hud_median_family_income	90300.0
census_tract_number	9901.0
application_date_indicator	2
action_taken_name	Preapproval request denied by financial instit...
applicant_income_000s	9999.0

dtype: object

Null Imputed Data Min Values

tract_to_msamd_income	14.05
population	5.0
minority_population	2.04
number_of_owner_occupied_units	10.0
number_of_1_to_4_family_units	10.0
loan_amount_000s	1
hud_median_family_income	48700.0
census_tract_number	1.0
application_date_indicator	0
action_taken_name	Application approved but not accepted
applicant_income_000s	-999.0

dtype: object

Binning

- Used Decision Tree Regressor to bin the features
- Passed the feature as independent and dependent variable in a decision tree regressor and the pruned the tree to bin the features.
- Playing with the parameters of the decision tree gives different amounts and types of bins

Thresholds

	tract_to_msamd_income	population	minority_population	number_of_owner_occupied_units	number_of_1_to_4_family_units	loan_amount_000s	hud_median_family_income	census_tract_number	applicant_income_000s
0	-1000.0	-1000.0	-1000.0	-1000.0	-1000.0	-1000.0	-1000.0	-1000.0	-1000.0
1	-2.0	-2.0	-2.0	-2.0	-2.0	-2.0	-2.0	-2.0	-2.0
2	81.98999786376950	3049.0	17.835000038147000	658.0	1323.5	197.5	58650.0	173.0	86.5
3	95.98500061035160	3989.5	24.28499984741210	956.5	1644.0	351.5	64450.0	364.5	155.5
4	109.9800033569340	5502.5	30.989999771118200	1202.5	1974.5	633.5	71100.0	611.5	284.5
5	124.625	6375.0	40.27000045776370	1433.5	2402.0	1345.5	84200.0	5064.510009765630	576.5
6	142.0199966430660	7338.5	51.41500091527300	1959.5	2958.5	100002.0	100002.0	9304.5	100002.0
7	170.6800003051760 0	8893.0	69.64500045776370	2338.5	4251.5			9564.0	
8	100002.0	100002.0	100002.0	100002.0	100002.0			100002.0	

Crosstable for feature tract_to_msamd_income

action_taken_name tract_to_msamd_income	Application denied by financial institution	Loan originated	% Loan originated
(-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

Crosstable For Feature Applicants Income

action_taken_name applicant_income_000s	Application denied by financial institution	Loan originated	% Loan originated
(-1000.0, -2.0]	4012	22495	0.848644
(-2.0, 86.5]	36874	108296	0.745994
(86.5, 155.5]	15943	87758	0.846260
(155.5, 284.5]	5559	35848	0.865747
(284.5, 576.5]	1301	7799	0.857033
(576.5, 100002.0]	488	1516	0.756487

- Here we see that more applicant income relates to more % of loan originated, but we will have to make lesser bins for the feature

Merging Bins of Cross Table Manually

action_taken_name applicant_income_000s	Application denied by financial institution	Loan originated	% Loan originated
(-1000.0, -2.0]	4012	22495	0.848644
(-2.0, 86.5]	36874	108296	0.745994
(86.5, 155.5]	15943	87758	0.846260
(155.5, 100002.0]	7348	45163	0.860067

Final Selected Features For Numerical Data

	tract_to_msamd_income	minority_population	loan_amount_000s	hud_median_family_income	applicant_income_000s	action_taken_name
0	(109.98, 124.625]	(-2.0, 40.27]	(197.5, 351.5]	(64450.0, 84200.0]	(86.5, 155.5]	Loan originated
1	(81.99, 95.985]	(-2.0, 40.27]	(197.5, 351.5]	(-2.0, 58650.0]	(-2.0, 86.5]	Loan originated
2	(81.99, 95.985]	(-2.0, 40.27]	(197.5, 351.5]	(64450.0, 84200.0]	(86.5, 155.5]	Loan originated
3	(142.02, 170.68]	(-2.0, 40.27]	(197.5, 351.5]	(64450.0, 84200.0]	(155.5, 100002.0]	Loan originated
4	(142.02, 170.68]	(-2.0, 40.27]	(351.5, 100002.0]	(64450.0, 84200.0]	(86.5, 155.5]	Loan originated
...	...	...	...	...	...	...
466561	(95.985, 109.98]	(-2.0, 40.27]	(-2.0, 197.5]	(64450.0, 84200.0]	(86.5, 155.5]	Preapproval request denied by financial instit...
466562	(95.985, 109.98]	(-2.0, 40.27]	(-2.0, 197.5]	(64450.0, 84200.0]	(-2.0, 86.5]	Preapproval request denied by financial instit...
466563	(95.985, 109.98]	(-2.0, 40.27]	(351.5, 100002.0]	(64450.0, 84200.0]	(-2.0, 86.5]	Preapproval request approved but not accepted
466564	(81.99, 95.985]	(40.27, 51.415]	(197.5, 351.5]	(64450.0, 84200.0]	(-2.0, 86.5]	Preapproval request approved but not accepted
466565	(124.625, 142.02]	(-2.0, 40.27]	(197.5, 351.5]	(84200.0, 100002.0]	(-2.0, 86.5]	Preapproval request approved but not accepted

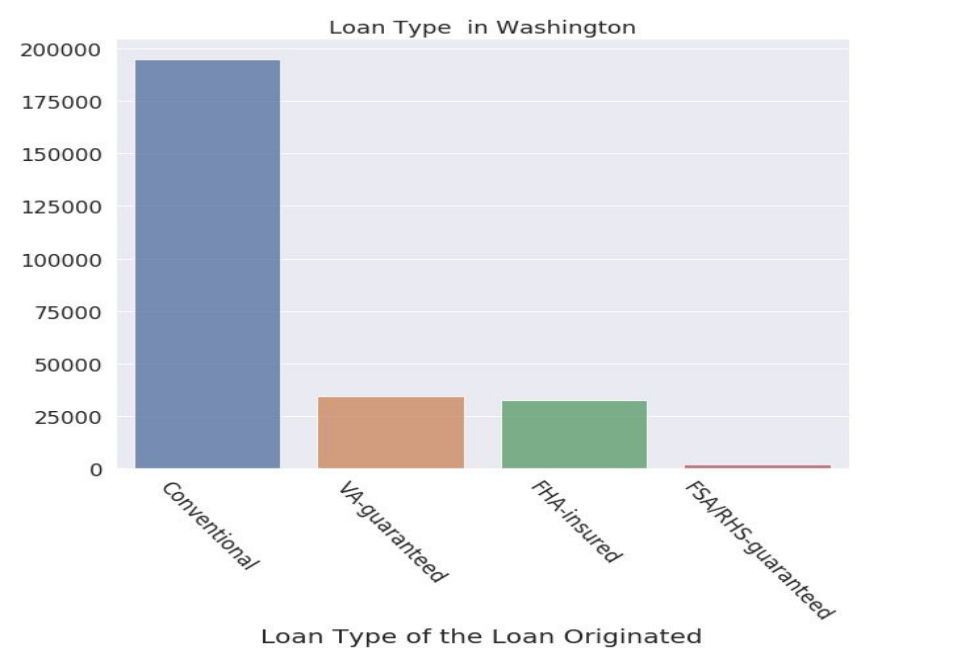
Categorical Data Features And NULL %

	0
purchaser_type_name	0.000000
property_type_name	0.000000
preapproval_name	0.000000
owner_occupancy_name	0.000000
msamd_name	8.203341
loan_type_name	0.000000
loan_purpose_name	0.000000
lien_status_name	0.000000
hoepa_status_name	0.000000
edit_status_name	84.031198
denial_reason_name_1	92.605762
county_name	0.078660
co_applicant_sex_name	0.000000
co_applicant_race_name_1	0.000000
co_applicant_ethnicity_name	0.000000
applicant_sex_name	0.000000
applicant_race_name_1	0.000000
applicant_ethnicity_name	0.000000
agency_name	0.000000
agency_abbr	0.000000
action_taken_name	0.000000

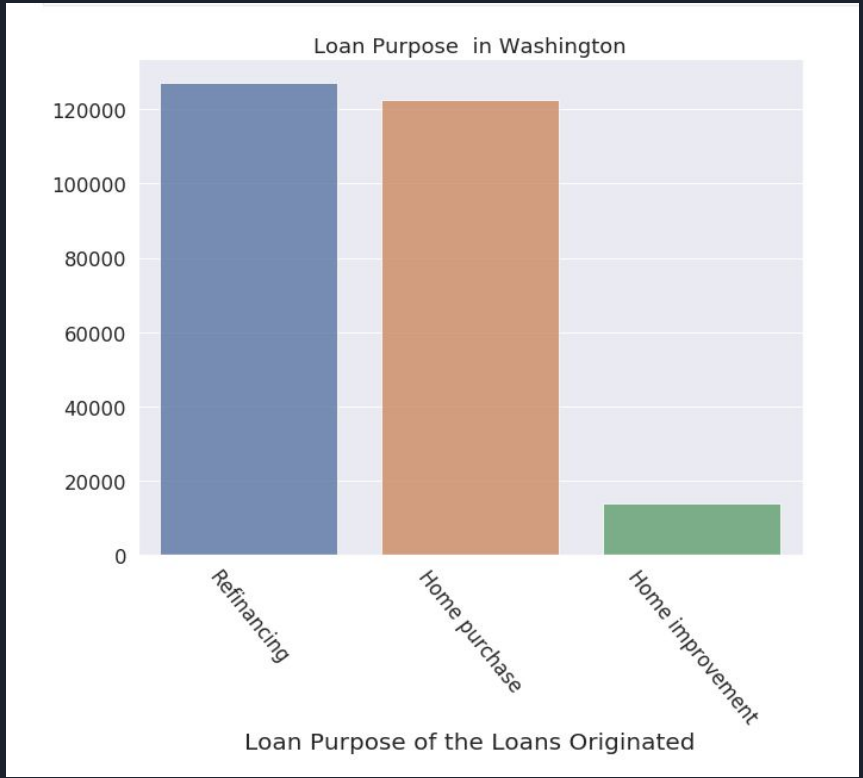
Handling of nulls in categorical features is done by creating nulls as a separate category. (as all of them have >5% nulls).

If some had <5% nulls, mode would have been used for imputation.

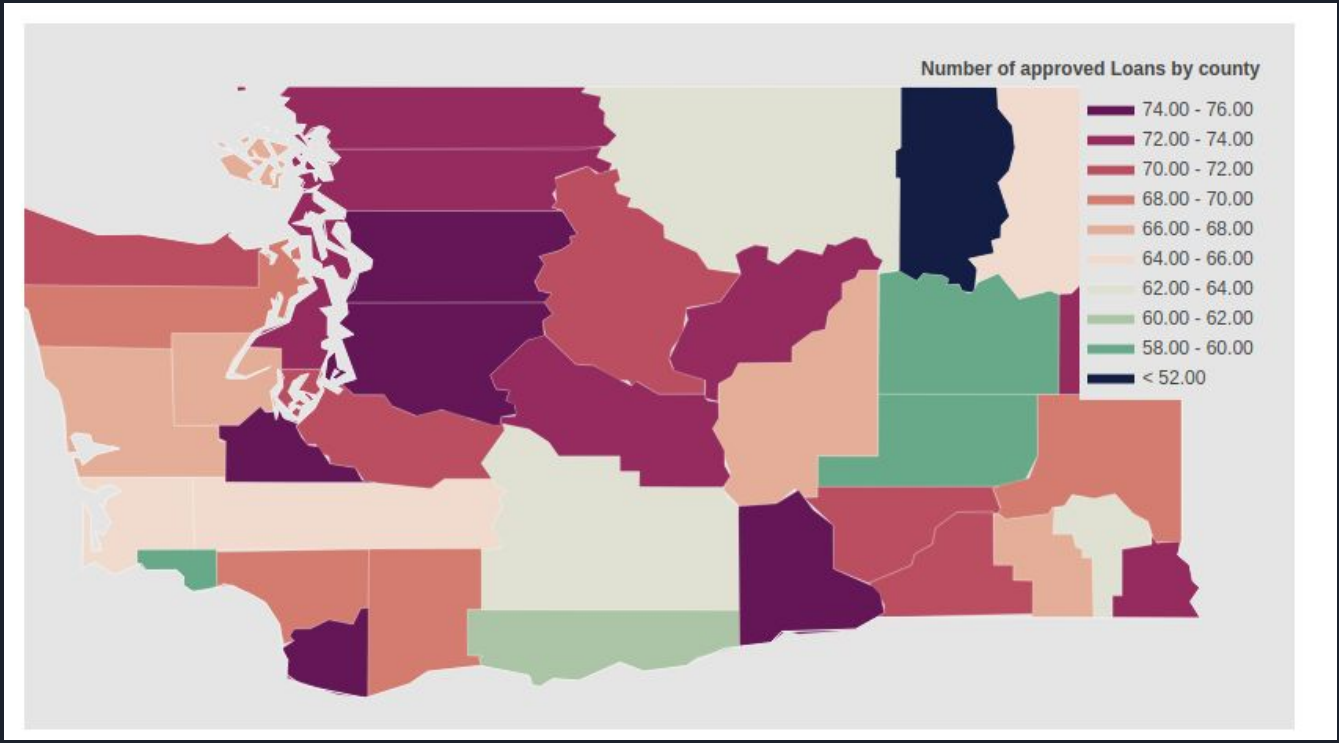
Categorical Features



Loan Purpose In Washington



Home Loans Approved By County



Cross Table of Categorical Feature applicant_sex_name

action_taken_name applicant_sex_name	Application denied by financial institution	Loan originated	% Loan originated
Female	17634	65579	0.788086
Information not provided by applicant in mail, Internet, or telephone application	6563	22433	0.773658
Male	39710	172650	0.813006
Not applicable	270	3050	0.918675

A few categorical features are also re binned to create less imbalance and more useful features.

Final Selected Features For Categorical Data

	action_taken_name	preapproval_name	loan_type_name	loan_purpose_name	co_applicant_sex_name	agency_name
0	Loan originated	Not applicable	Conventional	Others	Others	Consumer Financial Protection Bureau
1	Loan originated	Not applicable	FHA-insured	Home purchase	No co-applicant	Department of Housing and Urban Development
2	Loan originated	Not applicable	Conventional	Others	Others	Department of Housing and Urban Development
3	Loan originated	Not applicable	Conventional	Others	Others	National Credit Union Administration
4	Loan originated	Not applicable	Conventional	Others	Others	Federal Deposit Insurance Corporation
...	...	...	...	...	...	...
466561	Preapproval request denied by financial instit...	Others	FHA-insured	Home purchase	Others	Department of Housing and Urban Development
466562	Preapproval request denied by financial instit...	Others	Others	Home purchase	No co-applicant	Department of Housing and Urban Development
466563	Preapproval request approved but not accepted	Others	Conventional	Home purchase	No co-applicant	Department of Housing and Urban Development
466564	Preapproval request approved but not accepted	Others	FHA-insured	Home purchase	Others	Department of Housing and Urban Development
466565	Preapproval request approved but not accepted	Others	Conventional	Home purchase	No co-applicant	Department of Housing and Urban Development

466566 rows × 6 columns

Performing One-Hot Encoding

	tract_to_msamd_income_(81.99, 95.985]	tract_to_msamd_income_(95.985, 109.98]	tract_to_msamd_income_(109.98, 124.625]	tract_to_msamd_income_(124.625, 142.02]
0	0.0	0.0	1.0	0.0
1	1.0	0.0	0.0	0.0
2	1.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0
...	...	...	...	...
327884	0.0	1.0	0.0	0.0
327885	1.0	0.0	0.0	0.0
327886	1.0	0.0	0.0	0.0
327887	0.0	0.0	0.0	0.0
327888	1.0	0.0	0.0	0.0

327889 rows × 28 columns

Creating predictive models

1. Logistic Regression.
    a. Hyper parameters:
        i. penalty = 'l1'
        ii. solver = 'liblinear'
        iii. random_state = 100
        iv. C (inverse regularization parameter) = 0.5
    b. Results:
        i. AUROC: 71%

2. Gradient boosting classifier.

a. Hyper parameters:
    i. n_estimators = 200
    ii. learning_rate = 0.5
    iii. max_depth = 3
    iv. subsample = 0.7
    v. max_features = 0.7
    vi. random_state = 100
b. Results:
    i. AUROC: 73%

From the cross tables and model results, it can be concluded that, the dataset is good for exploration and visualization purpose but it doesn't have much predictive power.

Questions?

THANK YOU!