DAT102x: Predicting Mortgage Approvals from Government Data

Executive summary

Our goal is to analyze data and to design a model to predict whether a mortgage application was accepted (meaning the loan was originated) or denied according to the given dataset, which is adapted from the Federal Financial Institutions Examination Council's (FFIEC). Data contains variables about applicants, loan characteristics and amounts, location and population, etc. Basically, we have explored, analyzed data and looking for insights and relations between data. Then we have transformed it to a more power predictive form, if it is necessary, and build a machine learning model to predict our target variable. We can remark some conclusions and ideas:

- Presence of outliers (or data errors) in many numerical values.
- We can see that loans for home purchasing (loan purpose = 1) are most likely to be accepted than loans for home refinancing (loan purpose = 3)
- The loan acceptance is not affected by the loan type or the property type. No matter if it is a conventional loan or government-guaranted.
- Most of the applicants are white, not hispanic and male people and its ratio of acceptance is positive. But requests from black, hispanic people or women are slightly rejected.
- Lender is a categorical variable with too many different values, using a new variable related to the acceptance ratio of a lender is a much better option for predicting.
- Applications where loan amount is below 100,000 are more likely to be rejected. When it is higher than 150,000 they are more likely to be accepted.
- Applicants with incomes above 75,000 are more likely to be accepted.
- Applicants from locations where more than 60% of the population is from a minority slightly tends to be rejected. Even in areas with low percentage of minorities, applicants with relatively high incomes can be rejected

Finaly we would like to highlight that in this project analyzing, transforming and preparing the data had been the most relevant stage. The algorithm selected and its results are dependent of the quality of the data preprocessing step.

Data Exploration

Our first step is a brief exploration on the dataset provided, we want to get some rapid insights or ideas about our dataset. So first of all, we list all columns, datatypes, number of rows, …

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 500000 entries, 0 to 499999
Data columns (total 23 columns):
loan_type                          500000 non-null int64
property_type                      500000 non-null int64
loan_purpose                       500000 non-null int64
occupancy                          500000 non-null int64
loan_amount                        500000 non-null float64
preapproval                        500000 non-null int64
msa_md                             500000 non-null int64
state_code                         500000 non-null int64
county_code                        500000 non-null int64
applicant_ethnicity                500000 non-null int64
applicant_race                     500000 non-null int64
applicant_sex                      500000 non-null int64
applicant_income                   460052 non-null float64
population                         477535 non-null float64
minority_population_pct            477534 non-null float64
ffiecmedian_family_income          477560 non-null float64
tract_to_msa_md_income_pct         477486 non-null float64
number_of_owner-occupied_units     477435 non-null float64
number_of_1_to_4_family_units      477470 non-null float64
lender                             500000 non-null int64
co_applicant                       500000 non-null bool
accepted                           500000 non-null int64
row_id                             500000 non-null int64
dtypes: bool(1), float64(8), int64(14)
memory usage: 88.2 MB
```

So, we have 500,000 data records, grouped in 23 data columns, most of them numerical, some columns have missing values and one variable is boolean. Comparing the data with the problem description, we actually can define two groups of variables:

- Categorical (numerical but not a "number" or quantitative value): *loan_type, property_type, occupancy, preapproval, msa_md, state_code,county_code, applicant_ethnicity, applicant_race, applicant_sex, lender, co_applicant*
- Numerical: *loan_amount, population, minority_population_pct, ffiecmedian_family_income, number_of_owner-occupied_units, number_of_1_to_4_family_units,…*
- Label: accepted, this is our targeted variable

So, next we should inspect a brief descriptive summary of our dataset, showing the main statistics features. For the categorical variables:

	loan_type	property_type	loan_purpose	occupancy	preapproval	state_code	county_code	msa_md	applicant_ethnicity	applicant_race	applicant_sex	lender
count	500000.000000	500000.000000	500000.000000	500000.000000	500000.000000	500000.000000	500000.000000	500000.000000	500000.000000	500000.000000	500000.000000	500000.000000
mean	1.366276	1.047650	2.066810	1.109590	2.764722	23.726924	144.542062	181.606972	2.036228	4.786586	1.462374	3720.121344
std	0.690555	0.231404	0.948371	0.326092	0.543061	15.982768	100.243612	138.464169	0.511351	1.024927	0.677685	1838.313175
min	1.000000	1.000000	1.000000	1.000000	1.000000	-1.000000	-1.000000	-1.000000	1.000000	1.000000	1.000000	0.000000
25%	1.000000	1.000000	1.000000	1.000000	3.000000	6.000000	57.000000	25.000000	2.000000	5.000000	1.000000	2442.000000
50%	1.000000	1.000000	2.000000	1.000000	3.000000	26.000000	131.000000	192.000000	2.000000	5.000000	1.000000	3731.000000
75%	2.000000	1.000000	3.000000	1.000000	3.000000	37.000000	246.000000	314.000000	2.000000	5.000000	2.000000	5436.000000
max	4.000000	3.000000	3.000000	3.000000	3.000000	52.000000	324.000000	408.000000	4.000000	7.000000	4.000000	6508.000000

	loan_amount	applicant_income	population	ffiecmedian_family_income	number_of_owner-occupied_units	number_of_1_to_4_family_units	minority_population_pct	tract_to_msa_md_income_pct
count	500000.000000	460052.000000	477535.000000	477560.000000	477435.000000	477470.000000	477534.000000	477486.000000
mean	221.753158	102.389521	5416.833956	69235.603298	1427.718282	1886.147065	31.617310	91.832624
std	590.641648	153.534496	2728.144999	14810.058791	737.559511	914.123744	26.333938	14.210924
min	1.000000	1.000000	14.000000	17858.000000	4.000000	1.000000	0.534000	3.981000
25%	93.000000	47.000000	3744.000000	59731.000000	944.000000	1301.000000	10.700000	88.067250
50%	162.000000	74.000000	4975.000000	67526.000000	1327.000000	1753.000000	22.901000	100.000000
75%	266.000000	117.000000	6467.000000	75351.000000	1780.000000	2309.000000	46.020000	100.000000
max	100878.000000	10139.000000	37097.000000	125248.000000	8771.000000	13623.000000	100.000000	100.000000

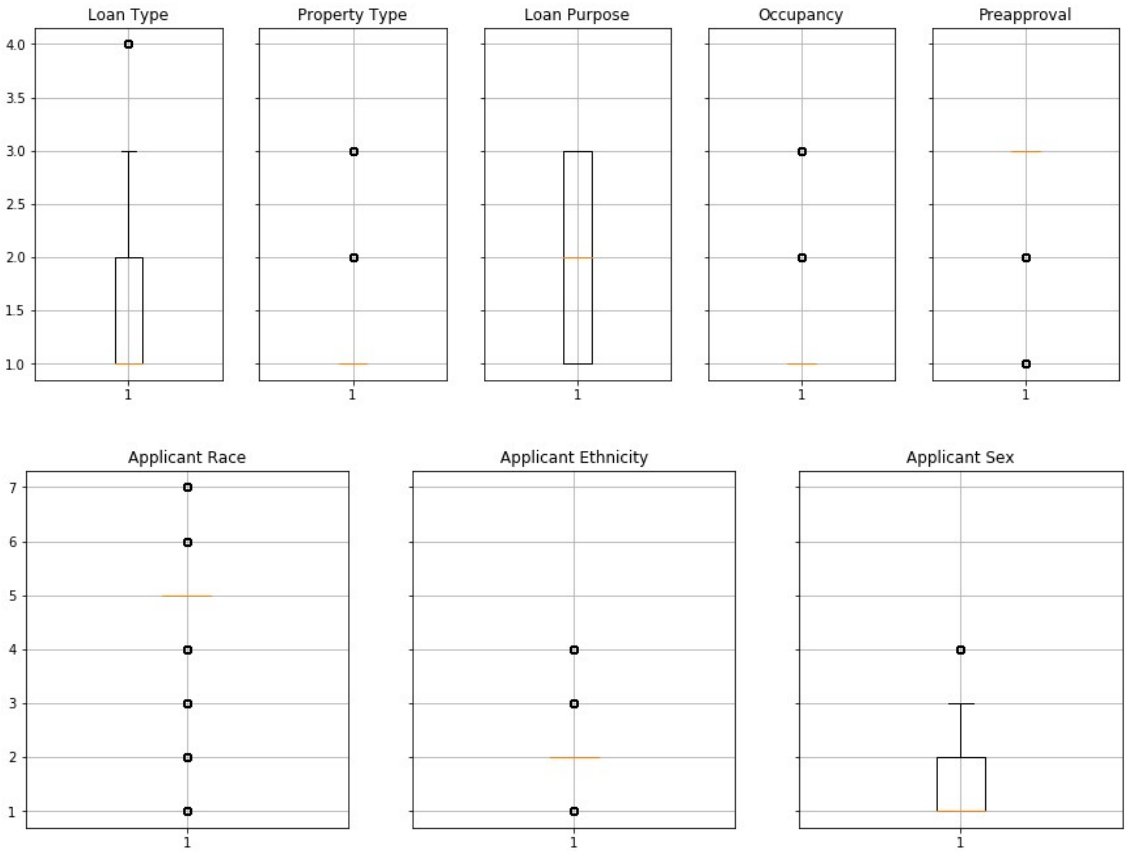
This brief data analysis give us some useful information:

- Many categorical variables take the same value for almost every row: propery_type, occupancy and preapproval.
- Approval value "It not applicable" for almost every row.
- Applicant race, ethnicity and sex have mainly two values, we will dive deeper later.
- There is no complete information about the location in many rows , and this should be an important variable.
- Presence of a lot of outliers (or data errors) in most of the numerical features. Especially loan amount and applicant income. Loan_amount mean is about 200 but maximum value is higher than 100,000 and standard deviation is almost 600. It looks like there are some wrongs values in the data.
- In general, the dataset seems to be balanced between accepted and not accepted applications

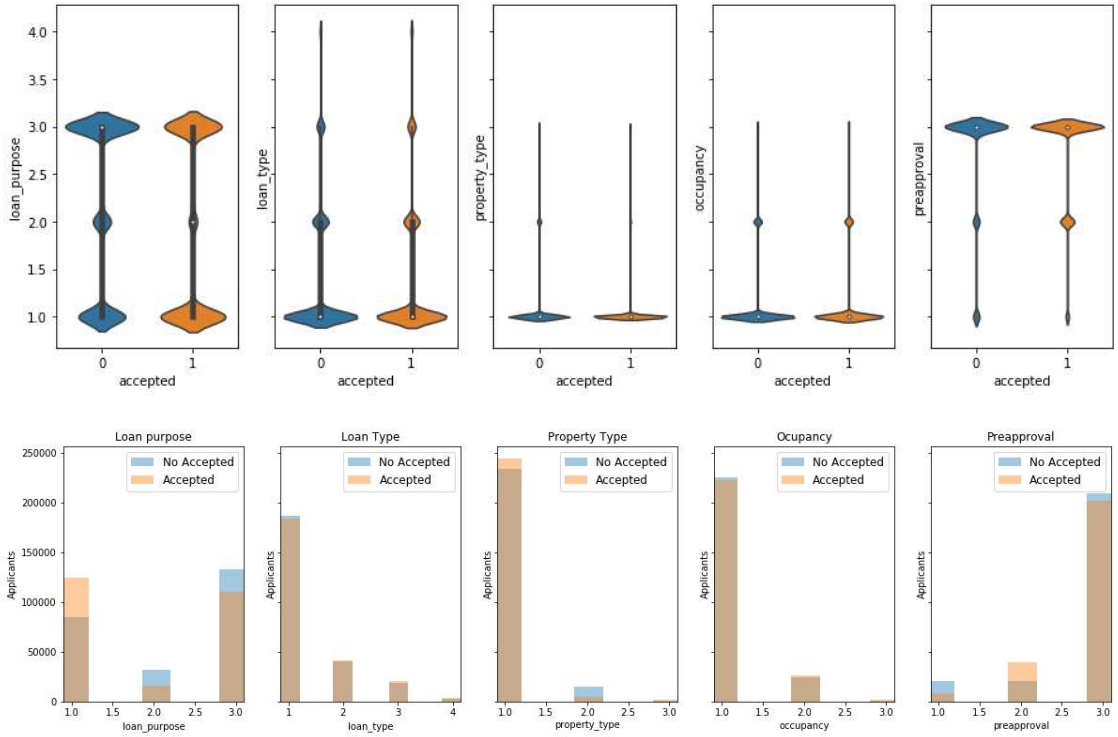
Let´s dive in every kind of data we have, searching for more useful information.

Exploring categorical data

First, a basic boxplot to identify how values are spread along its range



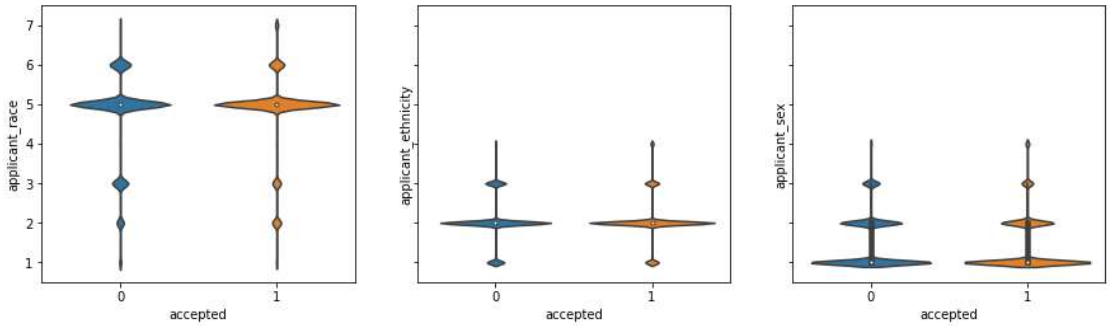
We use boxplot to see how values are spreaded along the range, many of the figures shows that most of the rows has a unique value and we should to explore if the label variable distribution is affected by some of these features. Let´s plot some graphics richer than the previous simple boxplot:

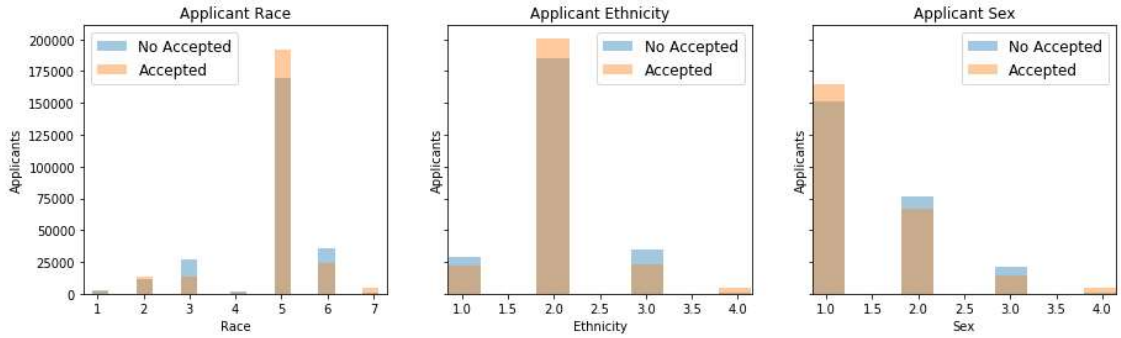


Now some ideas are shown:

- We can see that loans for home purchasing (loan purpose = 1) are most likely to be accepted than loans for home refinancing (loan purpose = 3).
- The loan acceptance is not affected by the loan type or the property type. No matter if it is a conventional loan or government-guaranted, they all have same opportunities. But most of applications are conventional loans (loan type = 1) and One to four -family properties (property type = 1)
- Owner´s principal dwelling (occupancy = 1) are the most frequent applications

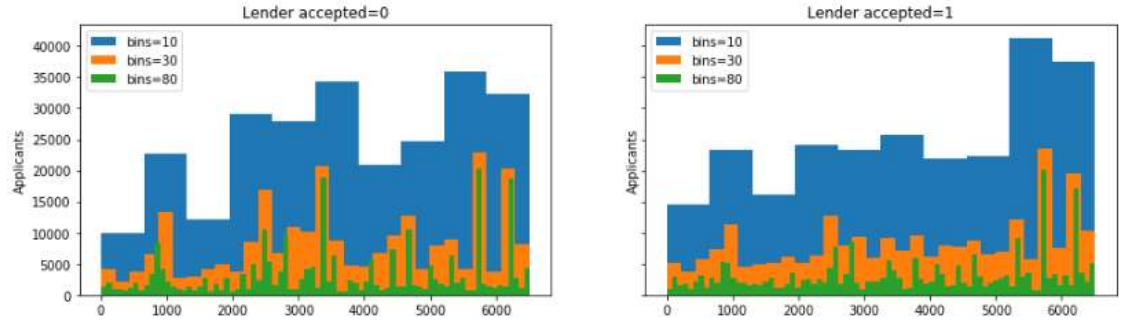
For features relatives to applicants we make the same analysis:





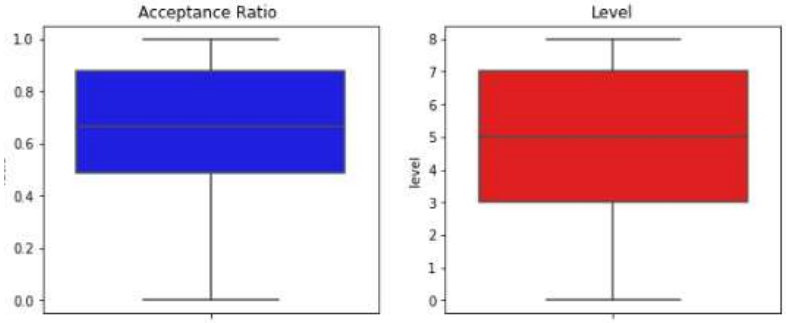
- Most of the applicants are white (5), not latino (2) and male (1) people and its ratio of acceptance is positive
- However, the requests of black (3), hispanic people (1) or women (2) are slightly rejected. But the difference is too small, we cannot confirm that it is a factor of discrimination.
- We appreciate that applicants who do not provide that kind of information tends to be not-accepted, so this information seems to be relevant for the lender.

Lender variable is an especial feature with a lot of different values as shown in the next figure:

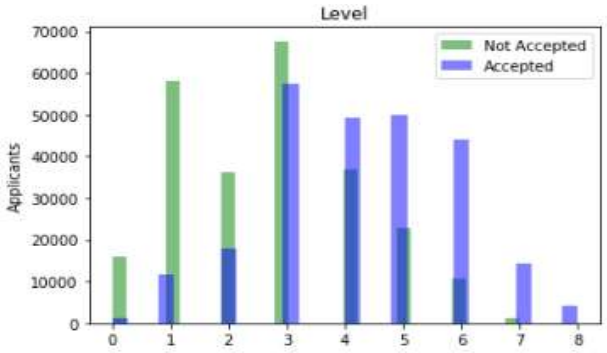


When we define only 10 bins, some kind of linear increasing tendency is shown. But when number of bins is increased the tendency is flatten, as it is supposed, but some peaks are revealed. We need to transform this data in order to get any sort of information.

We tried to consider the ratio of acceptance of a lender as a good piece of information, so for every lender, its acceptance ratio is calculated and we also need to reduce the number of categories: 6.1 thousand is not acceptable for a categoric variable. Then we decided to define levels of acceptance ratio: level 0 for 0%, level 1 for 0-12,5% ratio acceptance, level 2 for 12,5%-25% and so on.



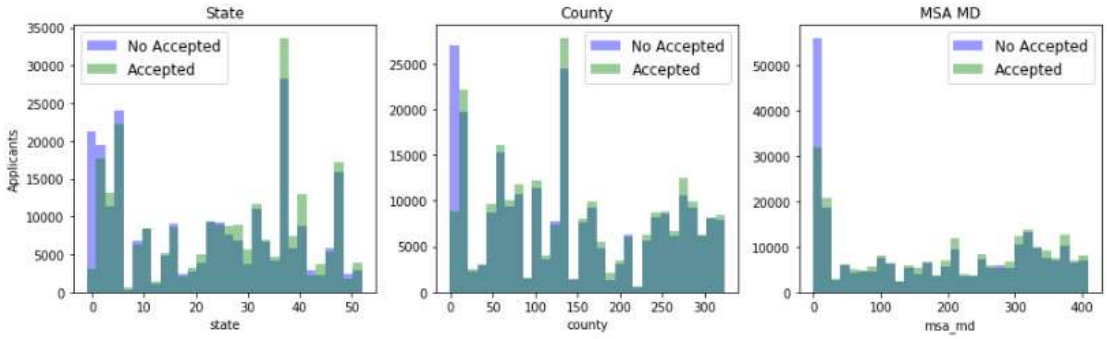
- The acceptance ratio for most of the lender is between 0.5 and 0.9, median 0.7 aprox. This mean there is some positive tendency in accepting loans.



The number of levels is not determined at this moment but 8 levels looks as a good option.

Location features

There are 3 variables related to the location, they all are categorical and exist too many values for them.

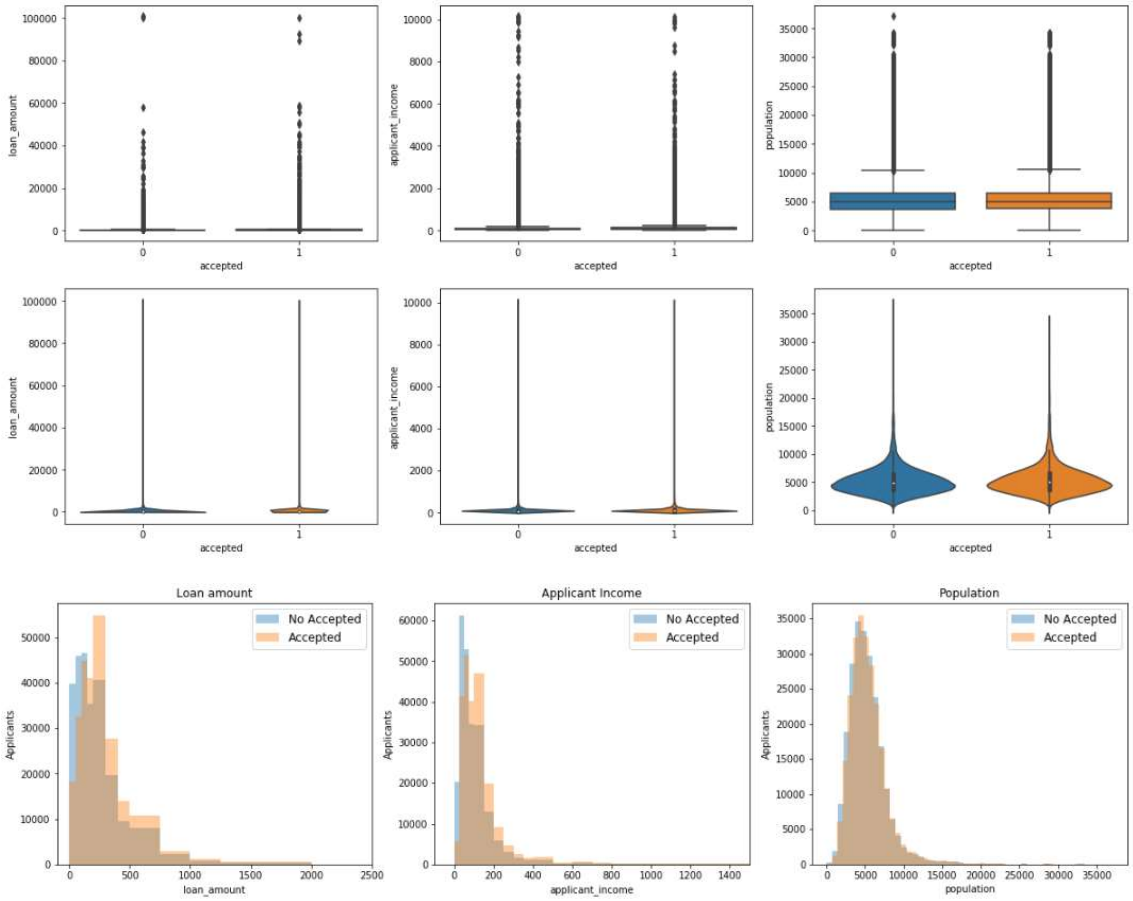


- There are a lot of records with a missing value (-1) in these columns and many of them are not accepted.
- Records with all of them as missing values (-1 in every column), we detect that there are missing values in some other numerical data so probably can not infer so much about them.

The state variable has values between 0 and 52 but there are no records with value 51, so its missing value could be replaced with the value 51. Maybe some kind of error occurred during the data processing or they are just missing. Applicants are equally distributed along all MSA MD values, so probably state is the best feature to represent the location.

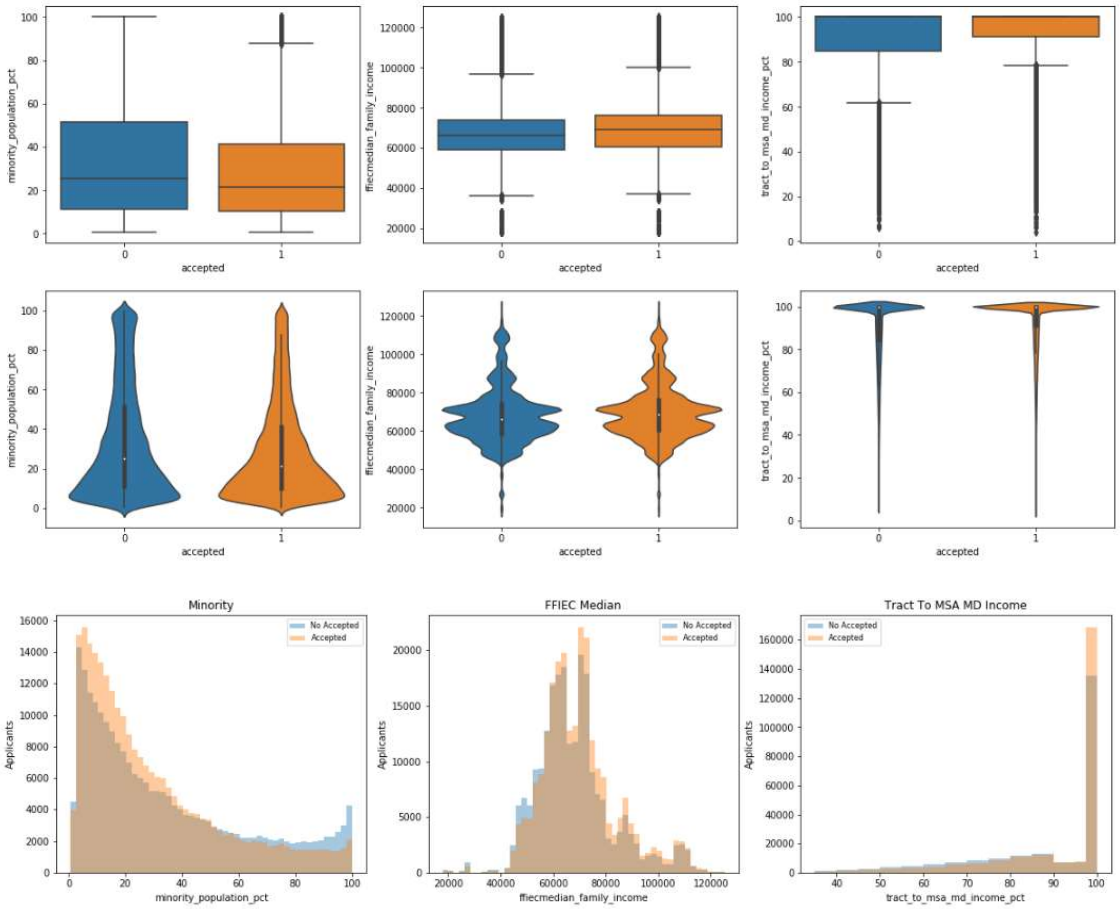
Exploring numerical variables

The next group of features to analyze are the numerical ones, we plot some graphs to show their distribution:



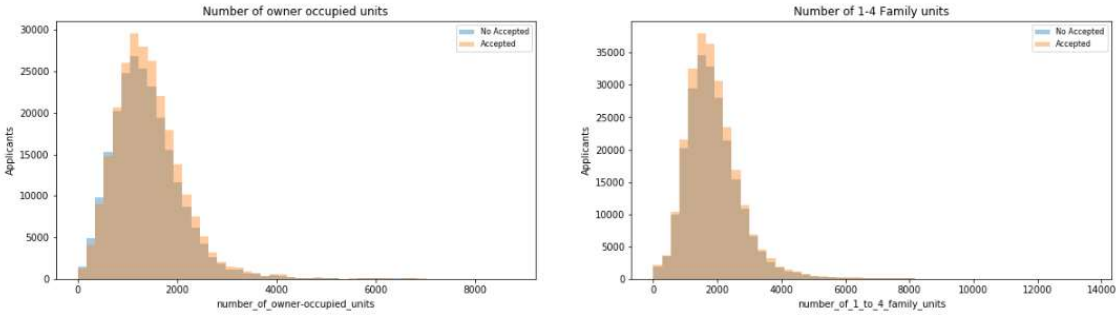
- We barely can see de interquartile range on the boxplot, so many outliers/errors are presents. There is no significance difference between accepted and not-accepted application distribution, even among the outliers.
- Those applications where loan amount is below 50,000 are likely to be rejected, between 50K-100K are some likely. When loan amount es higher than 150,000 the applications are more likely to be accepted.
- For applicant income, when applicant income is less than 75,000 are more likely to be rejected.

Repeat theses steps with the next group



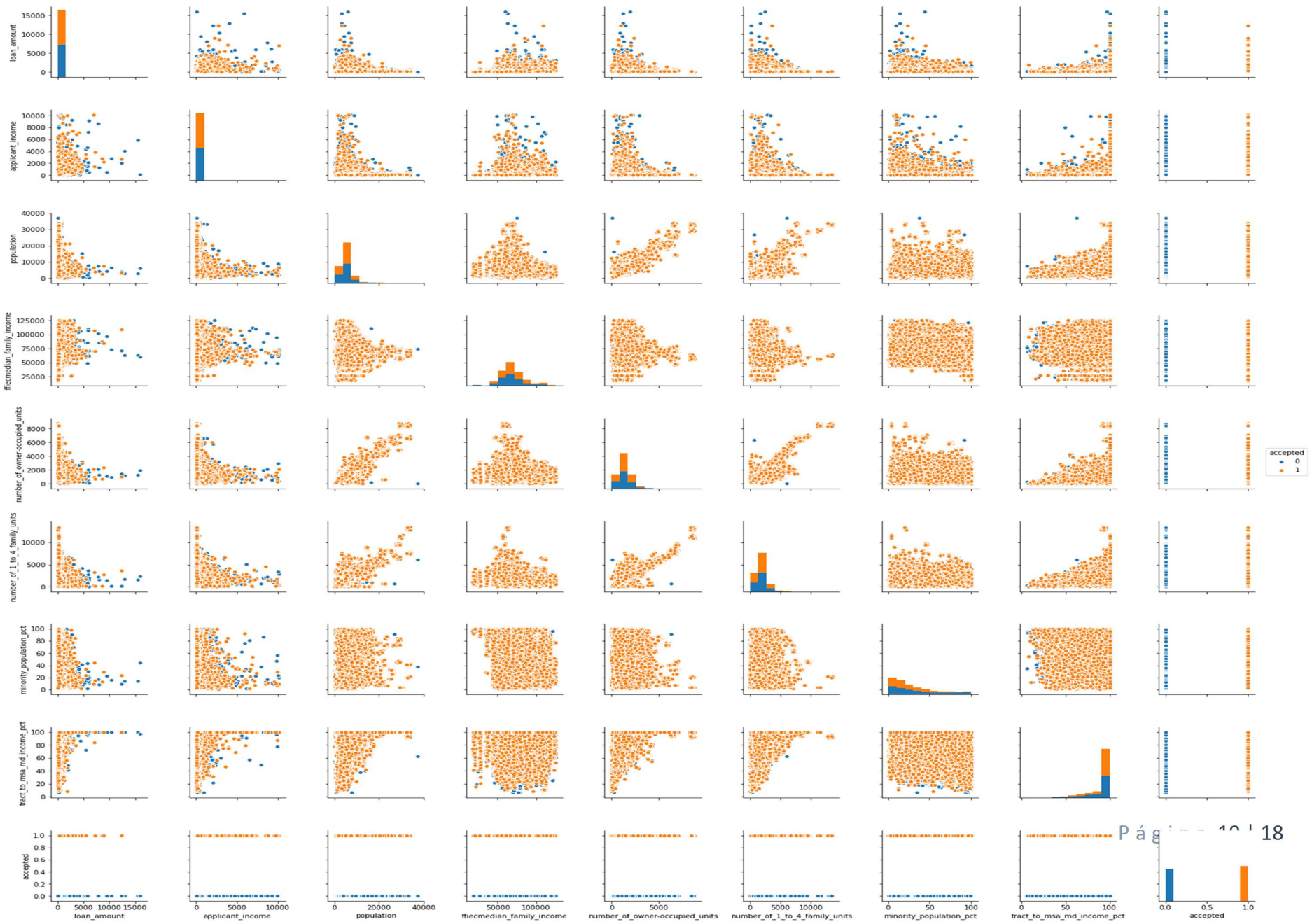
- Locations where the percentage of minority population is low, their applicants tends to be accepted. When 60% of population is from a minority the number of rejected applicants tends to get higher.
- As we detect previously, the lower the median of applicant income is the less likely to be accepted the loan is. Below 55,000$, the ratio of denied loans is bigger.

There are two features that present a normal distribution but we cannot appreciate any important remarkable point on them.



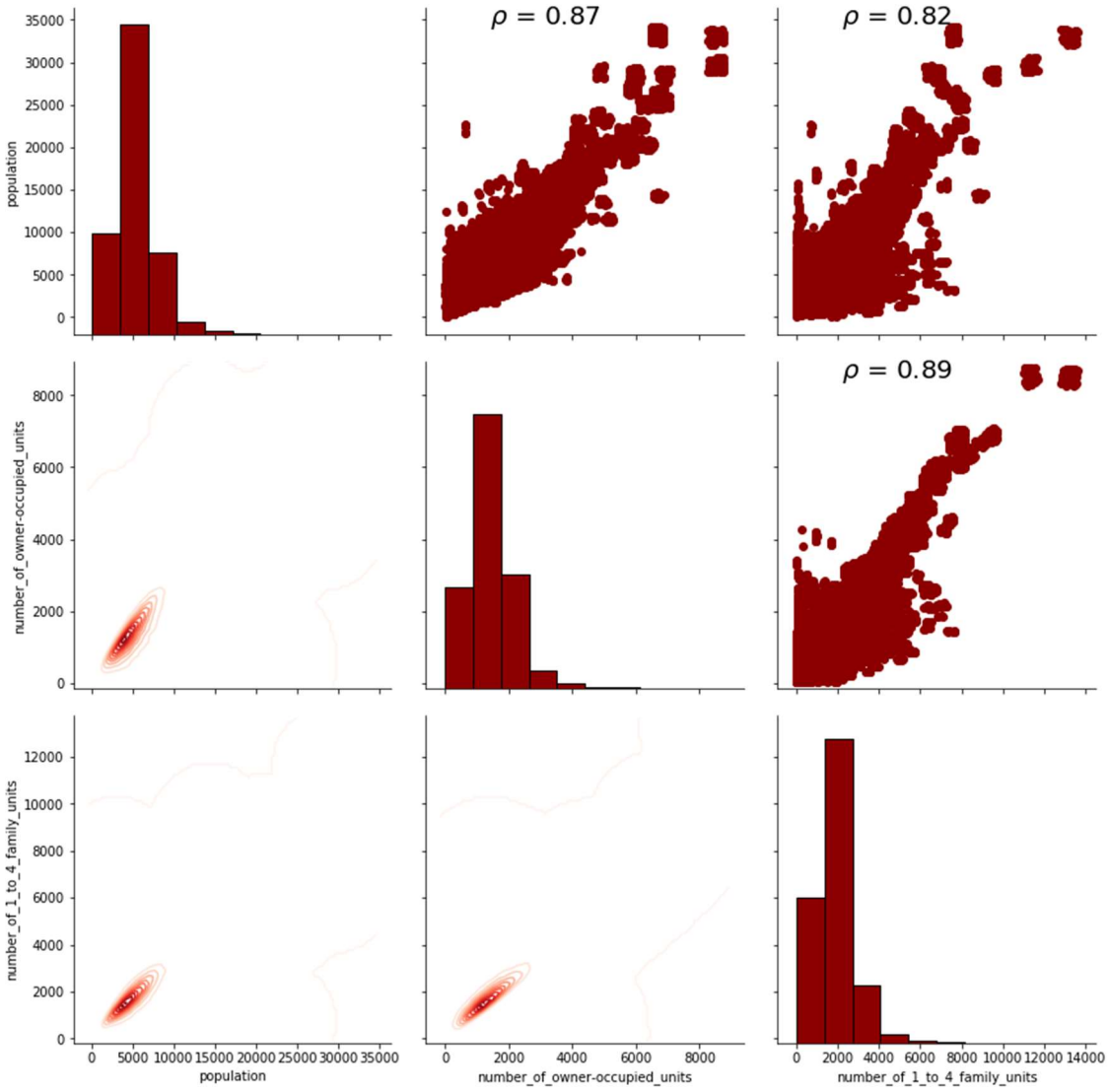
Analyzing relation between numerical variables

At this moment we have some knowledge about the features and how the acceptance ratio is dependent on that. Next step, analyze how these features are related between them. So a scatter matrix is our tool to get that insight:

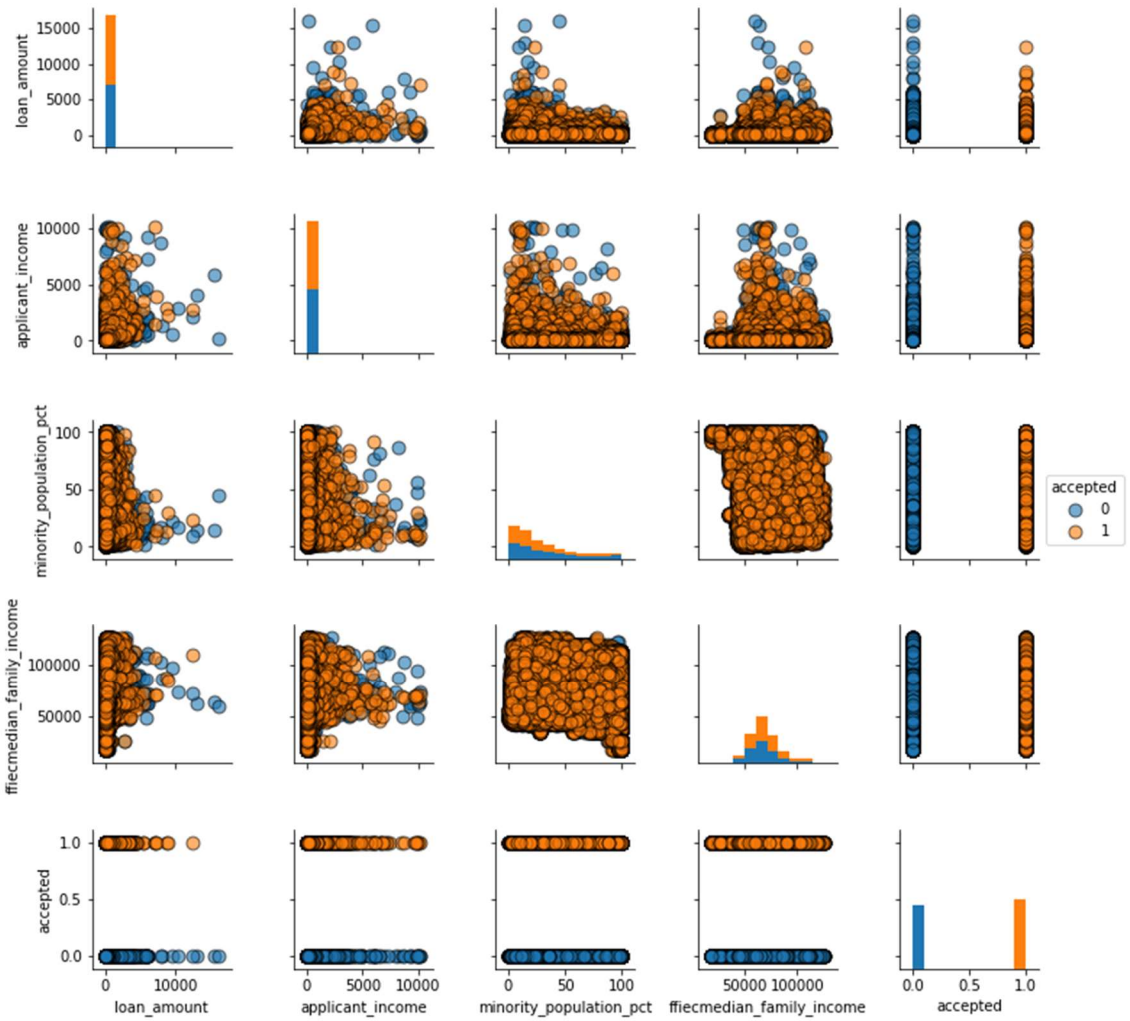

loan_amount
applicant_income
population
ffiecmedian_family_income
number_of_owner-occupied_units
number_of_1_to_4_family_units
minority_population_pct
tract_to_msa_md_income_pct
accepted
accepted
0
1

- The relationship between the variables loan amount or applicant incomes with respect to each of the remaining variables allows us to identify accepted and not accepted loans. Loan amount
- There is a strong linear correlation between population, number of 1-4 family units and number of owner occupied units.

	loan_amount	applicant_income	population	ffiecmedian_family_income	number_of_owner-occupied_units	number_of_1_to_4_family_units	minority_population_pct	tract_to_msa_md_income_pct	accepted
loan_amount	1.000000	0.482570	0.007791	0.275633	0.000216	-0.056754	-0.008875	0.165119	0.098280
applicant_income	0.482570	1.000000	-0.006849	0.115143	0.004775	-0.019644	-0.053969	0.102784	0.070343
population	0.007791	-0.006849	1.000000	-0.012020	0.860475	0.815828	0.089370	0.147899	0.020018
ffiecmedian_family_income	0.275633	0.115143	-0.012020	1.000000	-0.020078	-0.148316	0.020134	-0.049267	0.071361
number_of_owner-occupied_units	0.000216	0.004775	0.860475	-0.020078	1.000000	0.886270	-0.210577	0.355338	0.039244
number_of_1_to_4_family_units	-0.056754	-0.019644	0.815828	-0.148316	0.886270	1.000000	-0.155884	0.205387	0.006950
minority_population_pct	-0.008875	-0.053969	0.089370	0.020134	-0.210577	-0.155884	1.000000	-0.438936	-0.096945
tract_to_msa_md_income_pct	0.165119	0.102784	0.147899	-0.049267	0.355338	0.205387	-0.438936	1.000000	0.097640
accepted	0.098280	0.070343	0.020018	0.071361	0.039244	0.006950	-0.096945	0.097640	1.000000



We would lioke to focus in some features that seems to reveal some remarkable ideas:



- High values for loan amount and applicant incomes low are very likely to be rejected.
- There are no applications with a loan amount high in locations when % of minority population is very high, nor accepted or not accepted.
- Even in areas with low percentage of minorities, applicants with relatively high incomes are rejected

Classification model for acceptance of loan applications

The next point is to develop a predictive model that allows us to determine with an acceptable degree of confidence when a loan request or applicant is accepted or not. This is a two-class classification problem, accepted or not accepted. We will try to define a model starting from a simple approach. Then, this model will be refined but avoiding increasing complexity as much as possible.

After the analysis made we can take some decisions:

- Row_id variable, as it is well known, should be discarded because it does not provide any value or information
- The categorical variables loan type and accuracy should be removed
- Variables loan_property and approval do not seem to provide much value, although they are not completely disposable
- The variable lender has a number of values or categories too high and therefore it will be replaced by its level of acceptance ratio.
- About the variables related to race, ethnicity and sex of the applicant, some of them that are not predictive should be discarded, such as ethnicity or sex.
- Some of the numerical variables, those mentioned in previous sections, seem to be good candidates: loan_amount, applicant_income, minority population pct, applicant income or ffiec median family income
- Regarding the 3 variables related to the location of the applicants, we are goint to include only the state but others combinations will be analyzed.

Treatment of the variable lender

The variable lender as one would expect has an important value in the model but it presents a quantity of values, which become categories, very high which could be a source of overfitting. To address this problem we propose to transform this variable into a variable that defines ranges of acceptance ratio of a lender. Therefore for each lender we will calculate its acceptance ratio, as the number of accepted applications divided by the total of applications processed by that lender
Once these values are calculated, we will group the lenders in the ranges 0-20% of acceptance ratio, 21-40%, ... In this way we reduce the number of categories of the variable to only 5 or 6 values and also provide some useful information about the facility of a lender to grant a loan.

Dealing with location variables

We are referencing to the variables state, county and msa md (Metropolitan statistical area or metropolitan division), all of them define the location of the applicant and for all of them the number of values is very high. We also have a lot of records where one or more of these variables take the value -1 that indicate that its actual value is not known or has not been registered, that is, they are missing values.
We can also say that:

- For the state field we have values from 0 to 52 but the value 51 is missing. Which may lead us to consider that the value -1 in that variable is really the value 51.
- For the msa_md field the same thing happens, we have values from 0 to 408 but there is no record with the value 338, again it could be the real value of the records that take the value -1.
- In the case of county we have several values that never appear 85, ... so you can not make a simplification like the one mentioned in the previous fields

Finally, the initial approach will be to consider only the variable state as part of the predictive model and based on the results of our model we will consider to include the msa md field.

Dealing with outliers, data error and missing values

Again, we have null values in multiple of the numerical variables present in the problem and that can penalize the performance of the model that we are going to define. There are around 70 or 80 thousand records with missing values and also they appear

simultaneously in several variables at the same time. To approach this problem, we will try to fill in the null values of these variables looking for the value that can correspond to it based on other variables of the same registry:

- Applicant income: we will take the median value of the records belonging to the same state, county, msa_md, race, ethnicity and sex as the record with the null value. If it does not exist, we will search for that median value based on the state, the county, msa md, race and ethnicity and so on. Finaly if no median value exists we will take the mean value of that variable for the state of the record.
- Minority population pct: in this case when a record with a null value is found we will search the median value for the records of the same state, county or msa_md.If it does not exist it will be searched for the median of the same state and county or finally for the records of the same state.
- For the remaining variables, the approach used in the variable minority population pct will be used.

Regarding to outliers or possible data error, we have some variables that are seriously affected, such as loan amount, applicant income, ffiec median family income, etc. For the variables that are finally included in the predictive model, we will approach 2 methods:

- Those records that have values above the perceptible 98 will be removed from the training process
- Those records with values above the IQR * 1.5 will be replaced by the limit value defined by the IQR * 1.5

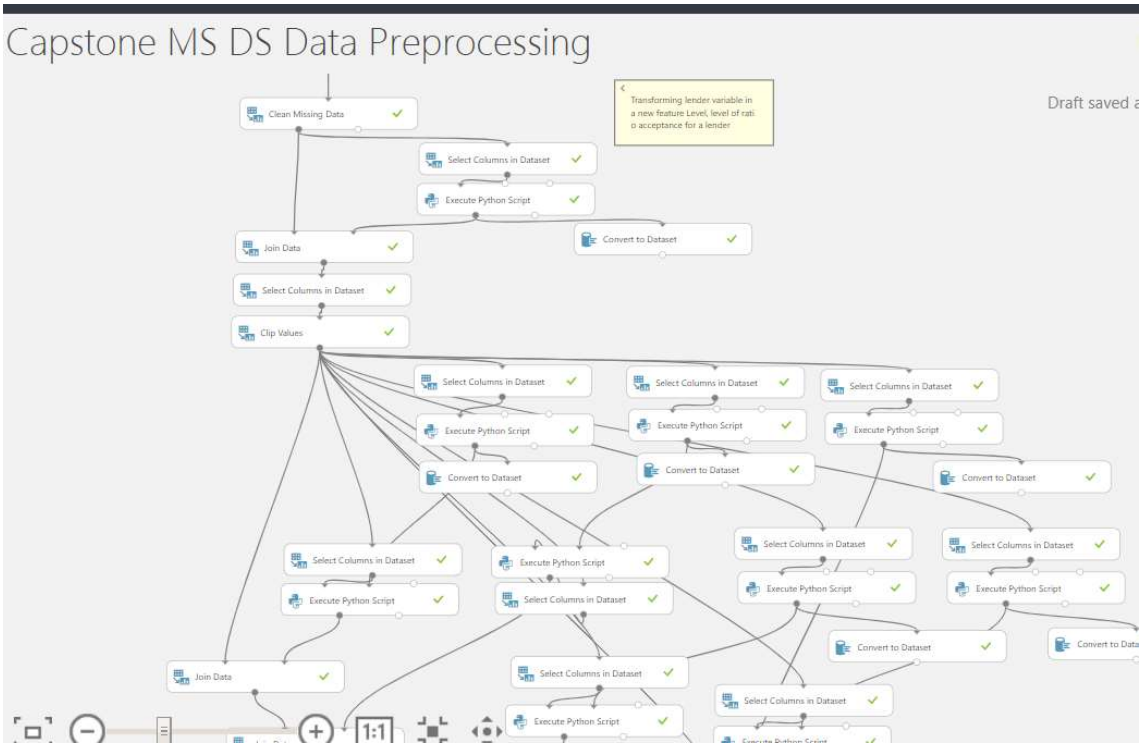
Building our predictive model

To build our model we will use Azure Machine Learning Studio cloud tool and some Python scripts to apply data transformations and manipulations. And we will do it in a group of stages: Training data preparation, predictive model design for training, test data preparation and scoring the model on the test data.

Data preparation

For this stage we design and Azure ML experiment where every data column is transformed base on the transformations defined previously:

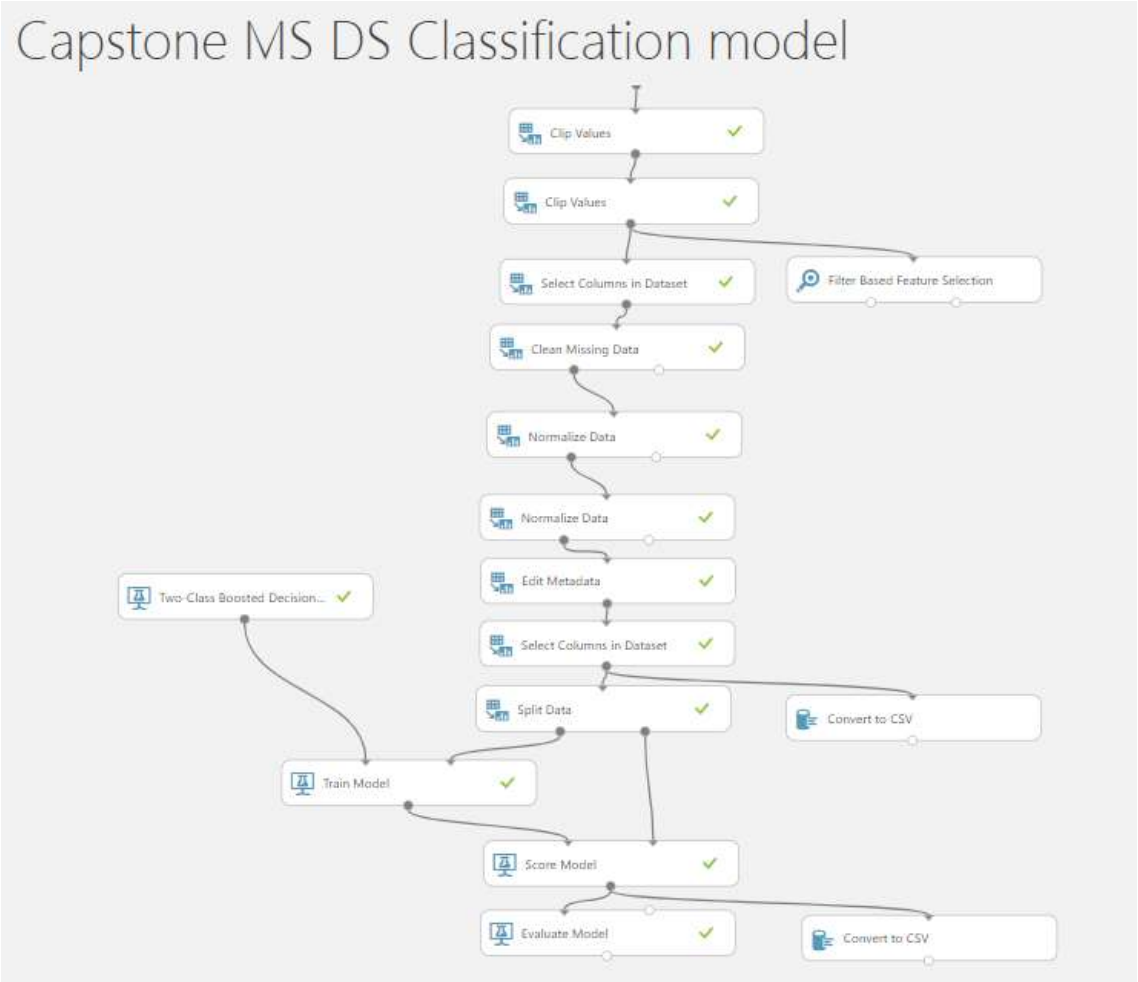
- transforming lender variable in an acceptance ratio level
- filling missing values for numerical variables based on the criteria mentioned previously.



So, this experiment produce a dataset that will be used for training our model and some others dataset to apply the same transformation to the test data.

Building our model

Our next experiment will receive the transformed dataset for training and apply a two-class classification algorithm to get a model for scoring the test data. Finaly we tried the main classification algorithm: logistic regression, decisión trees (and variations) and neural network. We evaluated every algorithm with the same data and multiples parameters and our best option was the Boosted Decision Tree algorithm.



We performed the following steps:

- clipping values above some threshold or percentile to deal with outliers
- remove some records with missing values
- Zscore normalization for some variables (loan_amount, applicant_income, ffiecmedian_family_income and population)
- Minmax normalization for minority_population_pct and tract_to_msa_md_income_pct)
- Make categorical variables for loan_purpose, level, applicant_race, etc…
- Selecting the columns to use in the training process
- Splitting the dataset in training set and test set: 75%-25%
- Train the Boosted decision tree
- Score and evaluate the results

After many experiments, the columns selected were: loan_amount, loan_pupose, applicant_income, applicant_race, state, minority population pct, tracto_to_msa_md_income_pct, ffiecmedian_family_income, ethnicity and lender level of acceptance ratio.
The training results were:

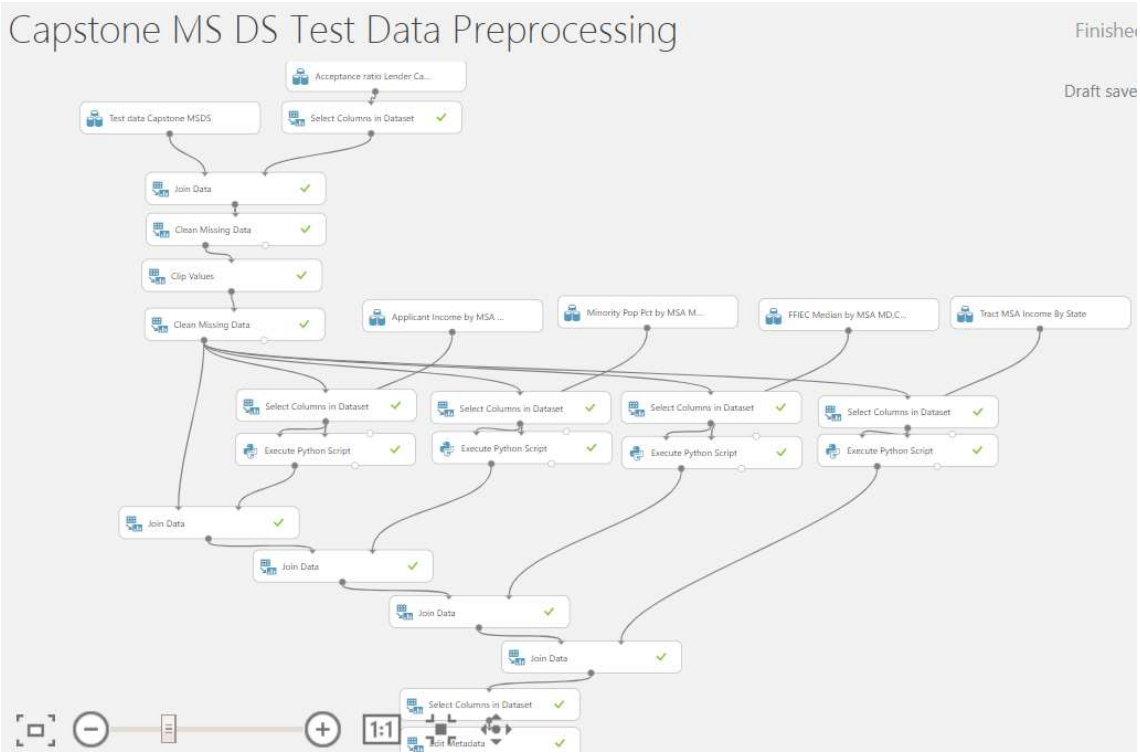
True Positive	False Negative	Accuracy	Precision	Threshold	AUC
38332	11703	0.721	0.703	0.5	0.799

False Positive	True Negative	Recall	F1 Score
16187	33778	0.766	0.733

Positive Label	Negative Label
1	0

Preparing the test data

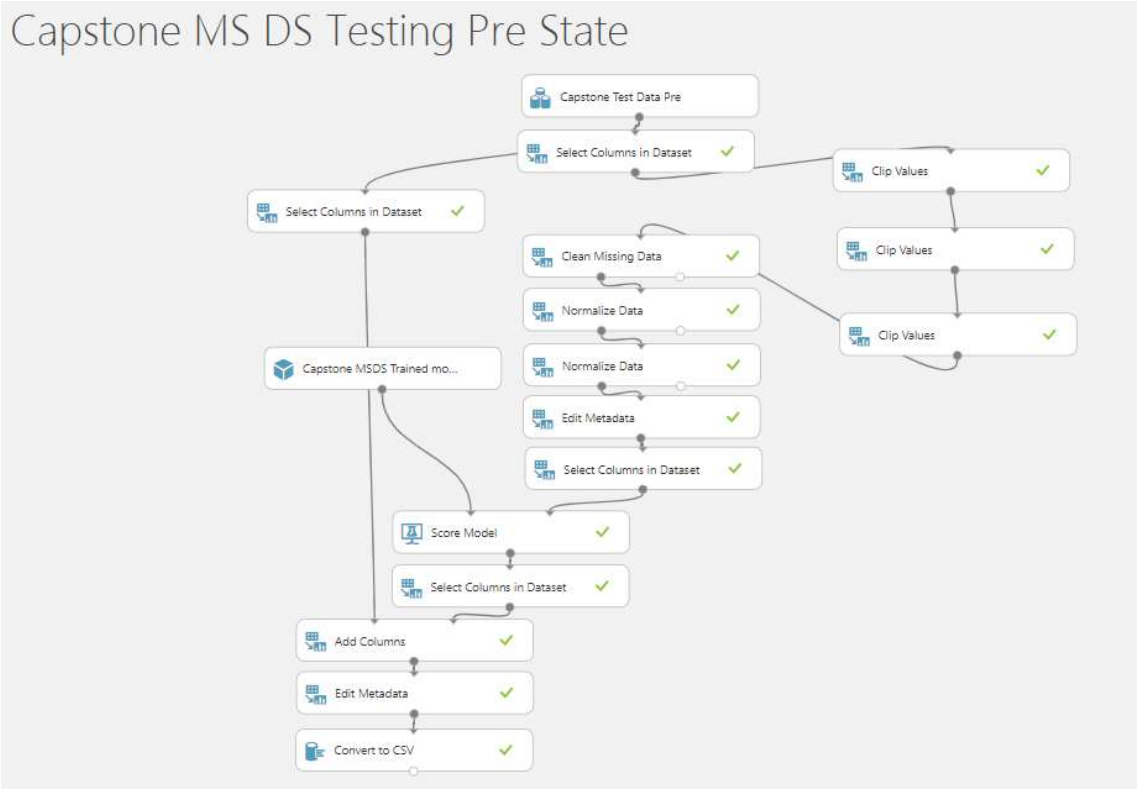
Now, we need to apply to our test dataset the same (or similar) transformations we have applied to the training data as we saw previously: transforming lender variable, filling missing values for some numerical variables (applicant_income, minority_population_pct, ffiecmedian, trac_to_msa_md)



The result of this process is a new dataset prepared to be used in our predictive model.

Scoring the model on test dataset

Finally we can apply our model to the test data and score every record to accepted or not accepted. This final experiment is:



The last step result in a csv file with the format required to submit the solution. Our final score was 0.7142 accuracy.

Conclusions

After all the time spent in analyzing and building the model we can conclude that data preparation has been the most powerful tool to get an acceptable performance. About the data analysis, we have yet mentioned many ideas: applicant race or ethnicity are relevant, applicants with high incomes as well as located in areas with low percentage of minority population are more likely to be accepted for a loan and most of the applications are related to the same type of property and loan.