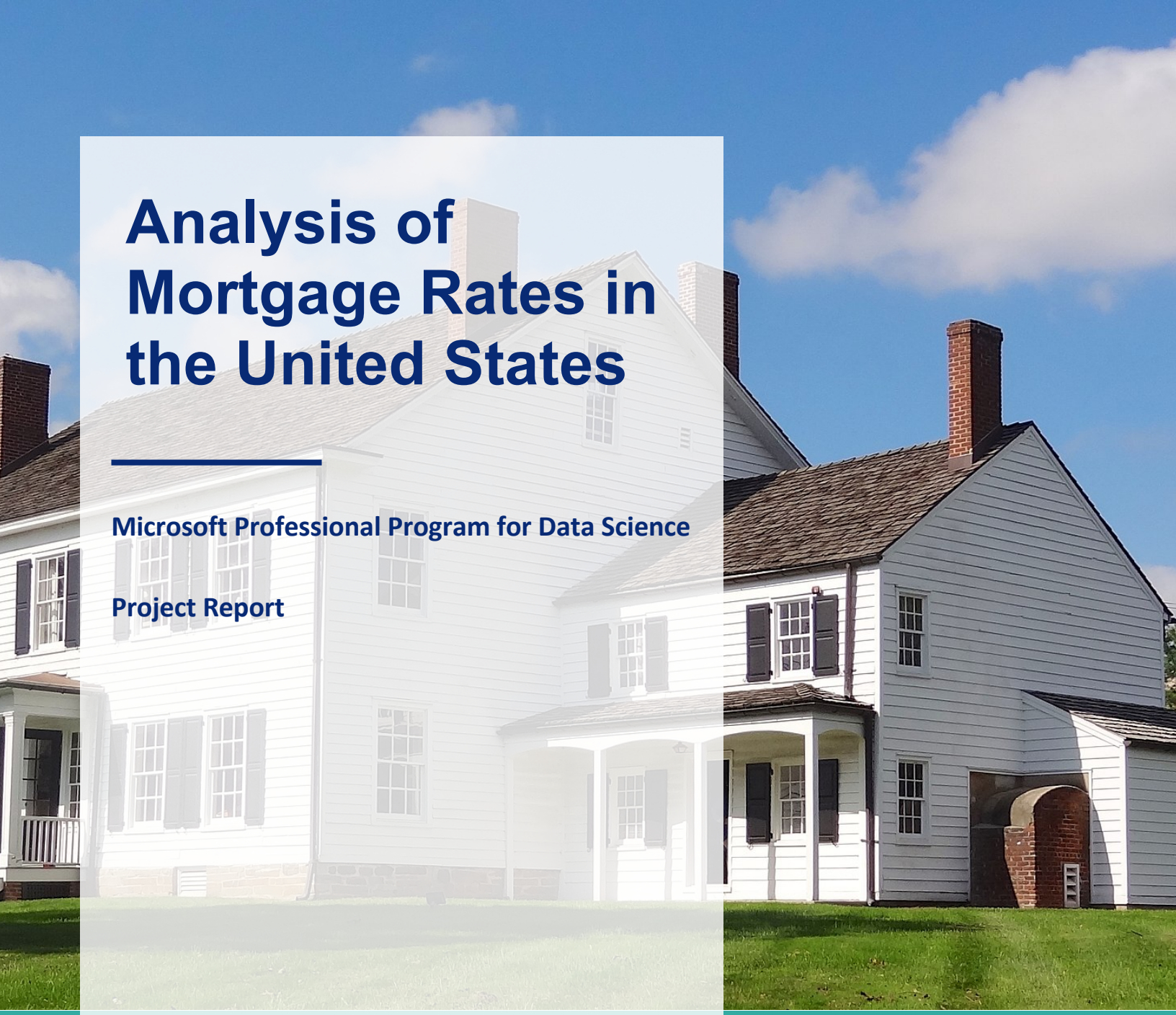


Analysis of Mortgage Rates in the United States

Microsoft Professional Program for Data Science

Project Report

NOVEMBER 25, **2019**

Authored by: Michael Ikemann

Analysis of Mortgage Rates in the US

Executive Summary

This document shall present an analysis of an adapted subset of the Home Mortgage Disclosure Act (HMDA) data, provided by the Federal Financial Institutions Examination Council (FFIEC). The HMDA was introduced by the US congress in 1975 and requires financial institutions like banks and credit unions to provide information about all loan applications they receive.

This transparency helps public officials to ensure lenders are serving the needs of the public and that the lending process follows fair rules and contains no discriminatory components. To support this verification mechanism the data includes next to required details such as income, the mortgage rate and the loan amount also detailed information about gender, state, county and city district, race and affiliation to a specific ethnic group of each applicant.

Our goal of this report is to verify that these requirements were fulfilled and to further support this with the insights won in the development process of a machine learning algorithm which can predict the mortgage rates of new applicants.

After a detailed analysis of the dataset and the evaluation of different machine learning approaches including neural networks and boosted regression trees, the author presents the following conclusions:

- **Fairness** – By analyzing the weightings of the single features the machine learning algorithm required for a highly accurate prediction of a mortgage rate there is no hint to any discriminatory influence on the rate provided. Neither could a relation be found between the affiliation to an ethnical group, an applicant's gender nor the race. The average mortgage rate of white people was though slightly higher than of black or Asian people - this could though in the further research be explained by a higher share of manufactured homes such as cheap caravans which

independent of the applicant's race usually have an overall much higher mortgage rate than e.g. an insured loan for a 1-4 family home.

- **Location** – There is a significant correlation between an applicant's state and city district and the mortgage rate provided and an even stronger correlation to specific counties.

- **Manufactured homes** – There is a very high average rate for manufactured homes of around 4% in comparison to an average of 1.5% for a classical house. It needs to be pointed out here though that applicants of loans for these houses usually also have a far lower average income of around $54,000 in comparison to an average of $78,000 of an applicant of a classical house.

- **Insured loans** – Loans insured by the Federal Loans Administration (FLA) which make up for around 55% of all the loans in the dataset as well as loans from the Veterans Administration and the Farm Service Agency and the Rural Housing Service have at average significantly lower rates of around 1.3% at average compared to an average of 2.5% for conventional, uninsured or otherwise subsidized loans.

- **Lender** – The overall largest direct influence on the mortgage rate has the lending institution. There was no data provided about each lender but after further analysis of the top 50 lenders using the data provided it could be shown that at least the huge majority of them are likely specialized on a specific type of loan. Many provide loans nearly exclusively for manufactured houses or FDA insured loans or both in combination. Thus, they effectively incorporate many of the other features into a new one with a very strong correlation between the average mortgage rate of their applicants.

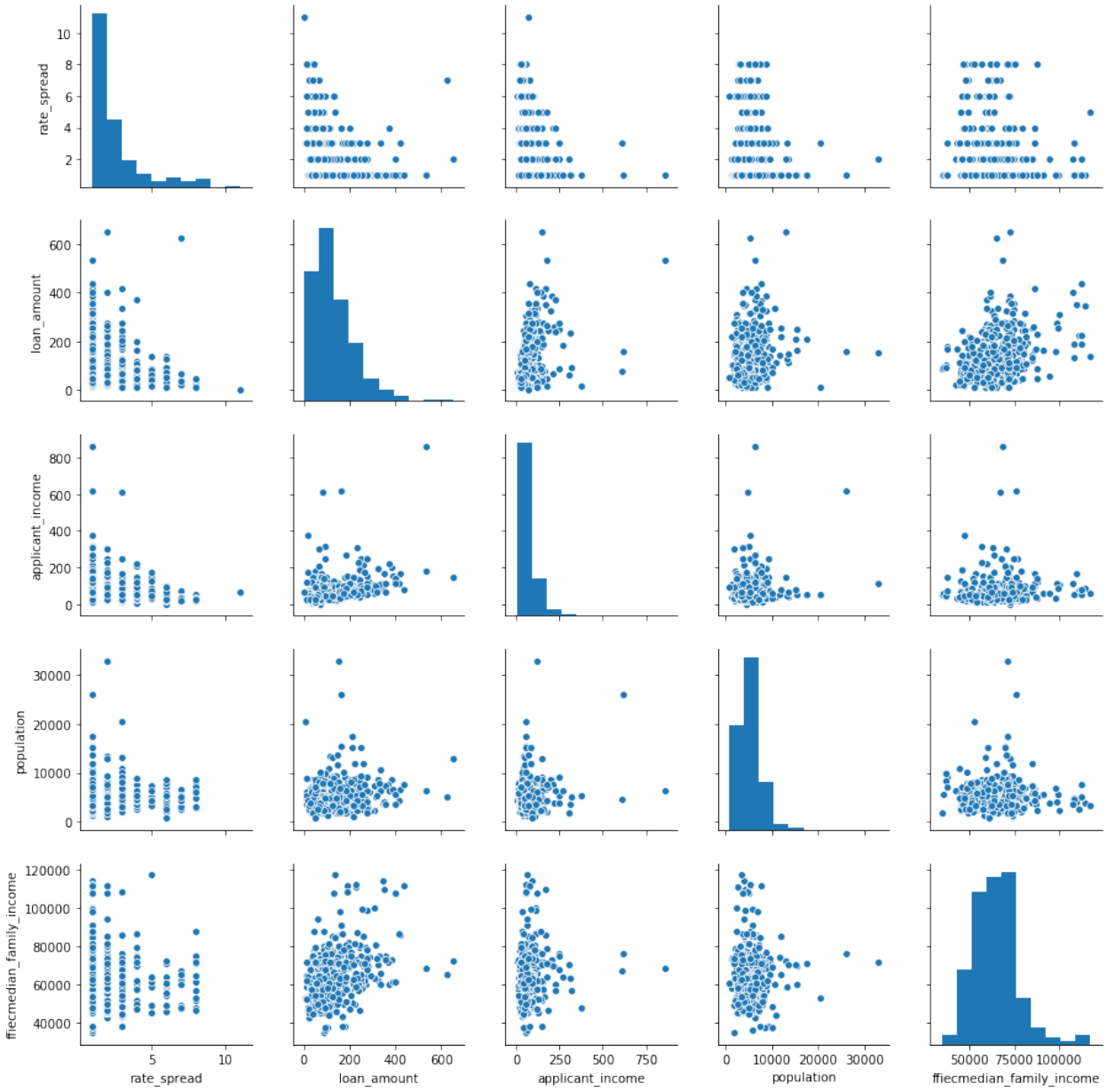
Initial Data Exploration

The initial data exploration began with a summary, descriptive statistics and the search for correlations between the provided features.

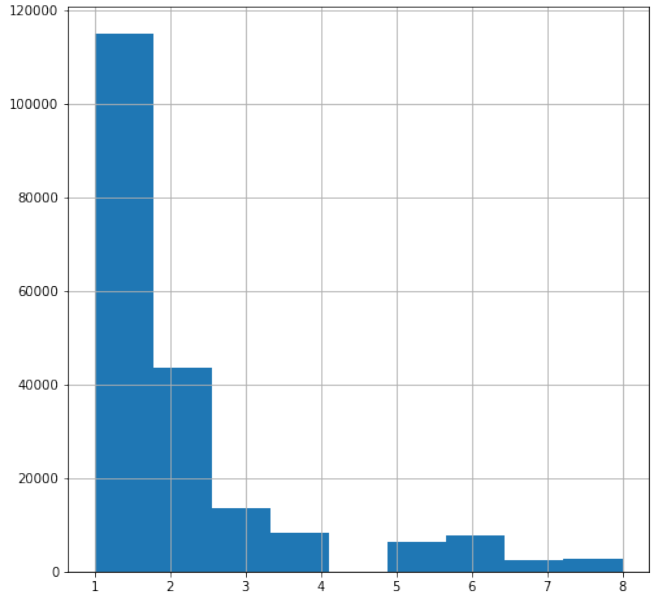
- The dataset consists of 400,000 entries of granted or preapproved mortgage loans, each entry consisting of 22 features further described below
- Around 6.5% of the entries were at least partially incomplete, around another 1% had serious outliers, e.g. mortgage rates of 99% p.a.
- The mean mortgage rate spread provided was at about 1.97% with a standard deviation of around 1.65%
- The average loan amount requested was $142,000 at an average income of $73,000. The maximum yearly income registered was as high as $10,000,000.
- There was just a minor correlation between incomes and a specific city district.

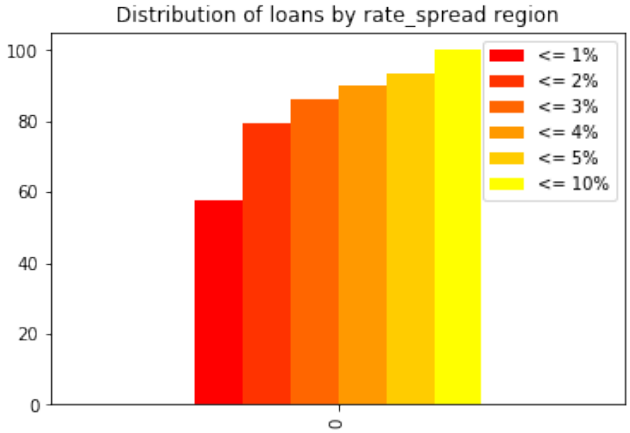
	rate_spread	loan_amount	applicant_income	population	minority_population_pct	ffiecmedian_family_income	number_of_owner-occupied_units
count	200000.000000	200000.000000	189292.000000	198005.000000	198005.000000	198015.000000	197988.000000
mean	1.979110	142.574940	73.617902	5391.099099	34.238640	64595.355801	1402.872401
std	1.656809	142.559487	105.696934	2669.028807	27.930882	12724.514485	706.880410
min	1.000000	1.000000	1.000000	7.000000	0.326000	17860.000000	3.000000
25%	1.000000	67.000000	39.000000	3717.000000	10.928000	56654.000000	932.000000
50%	1.000000	116.000000	56.000000	4959.000000	25.996000	63485.000000	1304.000000
75%	2.000000	179.000000	83.000000	6470.000000	52.000000	71238.000000	1742.000000
max	99.000000	11104.000000	10042.000000	34126.000000	100.000000	125095.000000	8747.000000

As you can see below the correlations between the single quantitative features were just minor except between the loan amount and the applicant's income – people with high incomes seem to buy more expensive homes – not surprising.



To get a better overview of the data I visualized the distribution of the loan rate spreads. It shows that the majority, so more than 80% are below or equal to a rate spread of 2% and even more than 90% of the loans provided were below 3%.





Individual Feature Statistics

The following features were provided in the data set:

General loan information

- **msa_md** – The Metropolitan Statistical Area/Metropolitan Division code (MSA/MD)
- **state_code** – A unique U.S. state identifier
- **county_code** – A ***per state*** unique county code
- **loan_amount** - The amount of money requested in thousands of dollars
- **loan_type** - Defines whether the loan granted/applied/purchased was conventional, government-guaranteed, or government-insured
    - 1 - Conventional (any loan other than FHA, VA, FSA, or RHS loans)
    - 2 - FHA-insured (Federal Housing Administration)
    - 3 - VA-guaranteed (Veterans Administration)
    - 4 - FSA/RHS (Farm Service Agency or Rural Housing Service)
- **property_type** - Defines for which amount of families the house was intended
    - 1 - One to four-family (other than manufactured housing)
    - 2 - Manufactured housing
    - 3 - Multifamily

- **loan_purpose** - Defines for which goal the loan was requested
    - 1 - Home purchase
    - 2 - Home improvement
    - 3 - Refinancing
- **occupancy** - Defines if the loan is intended for the owner's dwelling or otherwise
    - 1 - Owner-occupied as a principal dwelling
    - 2 - Not owner-occupied
    - 3 - Not applicable
- **preapproval** - Indicate whether the application or loan involved a request for a pre-approval
    - 1 - Owner-occupied as a principal dwelling
    - 2 - Not owner-occupied
    - 3 - Not applicable
- **lender** – A unique id for each lender of the loan

Applicant information

- **applicant_income** - The applicant's income in thousands of dollars
- **applicant_ethnicity** - Defines the applicant's ethnicity
    - 1 - Hispanic or Latino
    - 2 - Not Hispanic or Latino
    - 3 - Information not provided by applicant in mail, Internet, or telephone application
    - 4 or - Not applicable
- **applicant_race** - The applicant's ethnicity
    - 1 - American Indian or Alaska Native
    - 2 - Asian
    - 3 - Black or African American
    - 4 - Native Hawaiian or Other Pacific Islander
    - 5 - White
    - 6 - Information not provided by applicant in mail, Internet, or telephone application
    - 7 - Not applicable
    - 8 - Undefined
- **applicant_sex** - The applicant's sex
    - 1 - Male
    - 2 - Female
    - 3 - Information not provided by applicant in mail, Internet, or telephone application
    - 4 or 5 - Not applicable
- **co_applicant** - Defines if there was a co-applicant (like the applicant's spouse)

Census information – information about the city and neighborhood

- **population** – The total population of the living tract
- **minority_population_pct** – The percentage of a minority population to the total population for the tract
- **ffiecmedian_family_income** - FFIEC Median family income in dollars for the MSA/MD in which the tract is located (adjusted annually by FFIEC)
- **tract_to_msa_md_income_pc**t - Percentage of tract median family income compared to MSA/MD median family income
- **number_of_owner-occupied_units** - Number of dwellings, including individual condominiums, that occupied by the owner
- **number_of_1_to_4_family_units** - Dwellings that are built to house fewer than 5 families

For the training data (200,000 rows / 50% of the data):

- **rate_spread** – The difference between the effectively granted loan and the current standard rate.

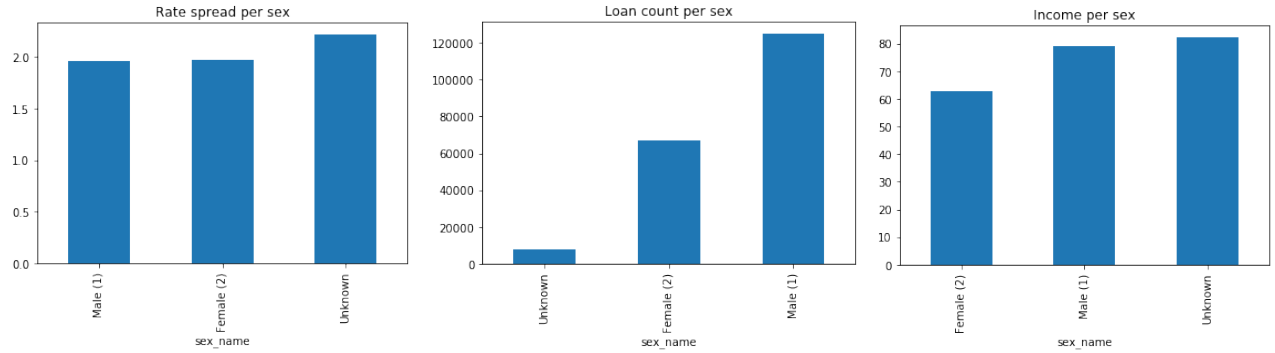
Correlations and Apparent Relationships

To find correlations between specific categorical features and the loan rate spread granted I first created a set of 30 features to binary/one hot encode categories such as male, female, race or ethnicity into a single feature per sub-category.

As one of the major goals of the HMDA is to prevent unfair and discriminatory loan rates I initially created an overview of the relation between these features and the rate_spread as well as a general overview of these features.

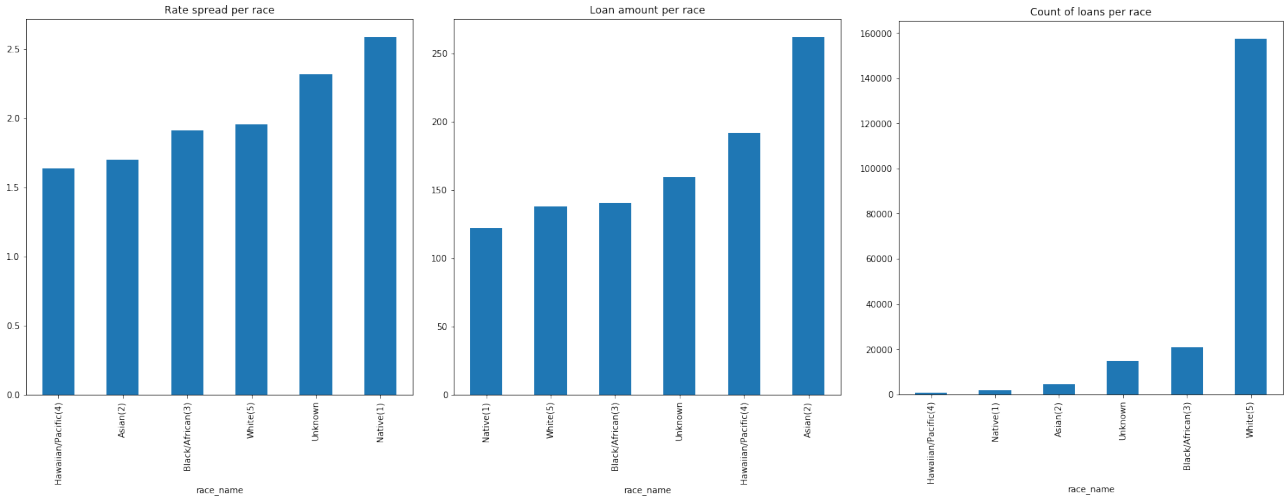
Gender

There is no significant difference between the loan rate offered to a male or female applicant. The rate is though slightly increased for applicants not willing to provide their gender at all. What we can also extract from the information provided is the fact that in around 60% of the cases a man requests the loan and just in around 35% a woman, though in the first case there is very often a co-applicant. As well we can see that a female applicant's average income is about $62,000 p.a. whereas the average income of man was $79,000, so significantly higher.



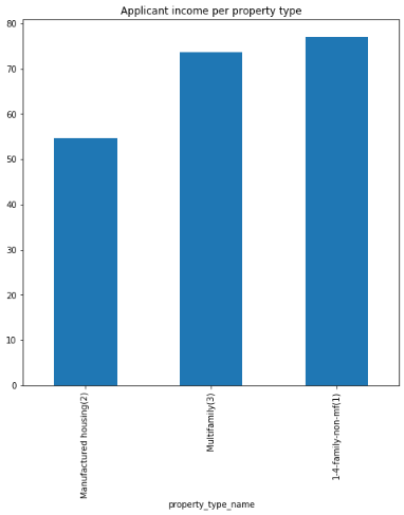
Verifying fairness due to race and ethnical affinity

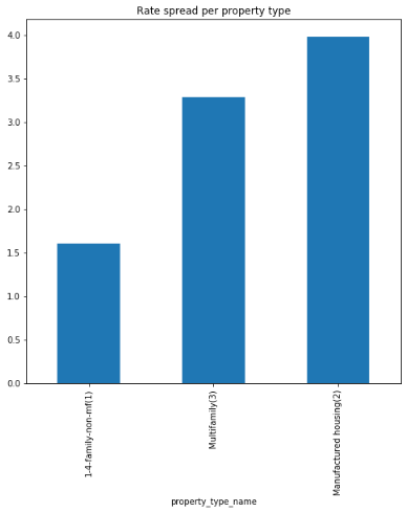
In the next step I analyzed if the loan rate is also independent of race or ethnicity. After starting with the gender you can see the distribution by race and the average loan amount below:

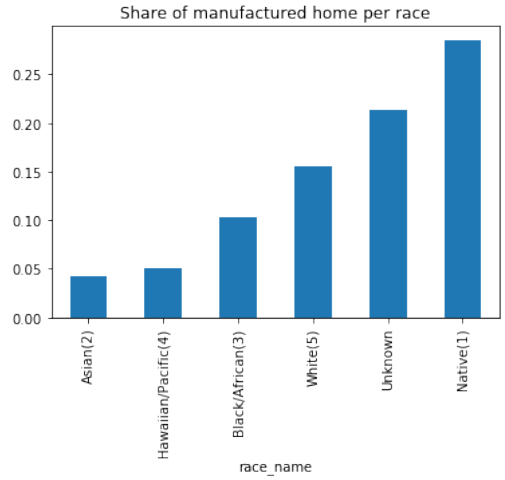


As you can see there is a slightly uneven distribution per race. The mean rate is a bit higher for white people than for black people, Asian people and Hawaiian people. So are loan institutes discriminating white people? No, likely not. Actually it turns out, we will still see this in detail in the later analysis, that white people - and the central graph showing the loan amount per race is already a hint to this – are far more often also totally fine with a “manufactured home” whereas Asian and Hawaiian people tend to either buy an expensive house or none at all. As we will still see the non-manufactured houses though have usually a far lower mortgage rate. In addition, people who buy a

manufactured house as shown below also usually have a lower income. To sum it up the higher average mortgage rate for white people and lower rate for Asian people is with high likeliness caused by their preference and acceptance of the home types.







Numeric Relationships

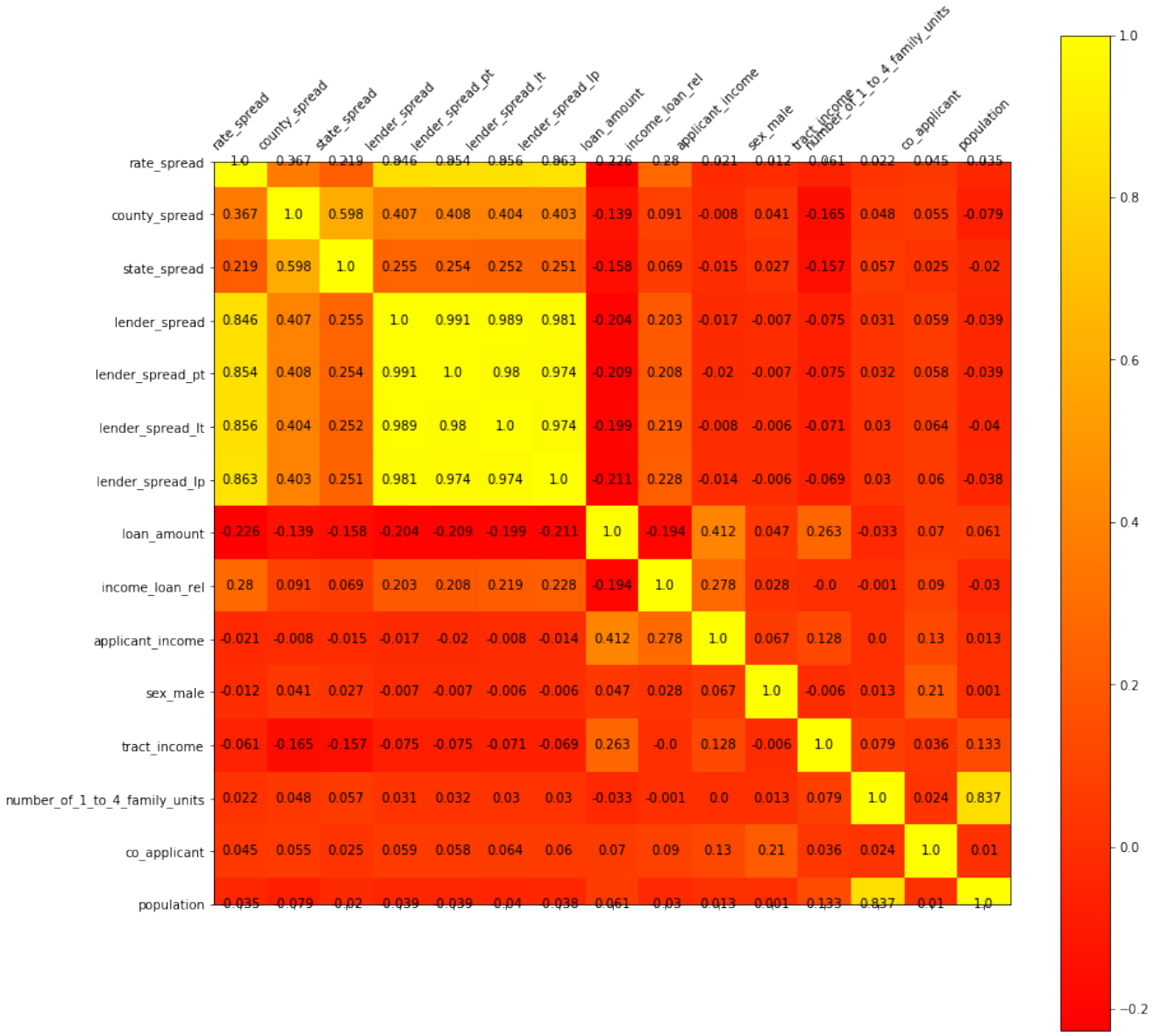
In the next step I analyzed if there are quantitative features with a direct correlation to the mortgage rate spread we shall predict with a machine learning algorithm. There were none which had a huge correlation in their original form.
What was quite striking though was a very strong variance from state and to state and even more from county to county. Further the credit type and its purpose seem to have a huge influence on the final rate. Another influence factor is the relation between the loans amount and the applicant's income. The bigger the discrepancy here the higher the chance of a default.

I engineered the following new features:

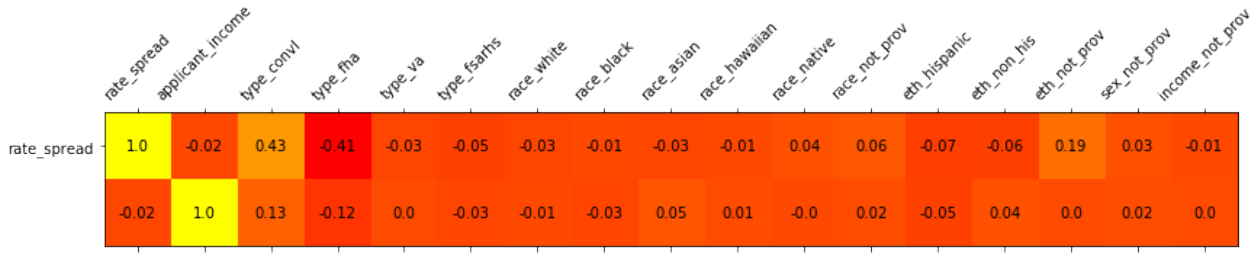
- The average loan spread per state – state_spread
    - The average loan spread per county in this state – county_spread
        - The spread per MSA in the county – msa_spread
- The average spread of the loan type (e.g. insured or not)
- The average spread of the purpose type (e.g. caravan or house)

- The average spread of the lending institute – lender_spread
    - Grouped by loan type as lender_spread_lt
    - Grouped by loan purpose as lender_spread_lp
    - Grouped by property-type as lender_spread_pt
- The relation between the applicant's income and the loan amount as income_loan_rel
- Binary features for all categorical features

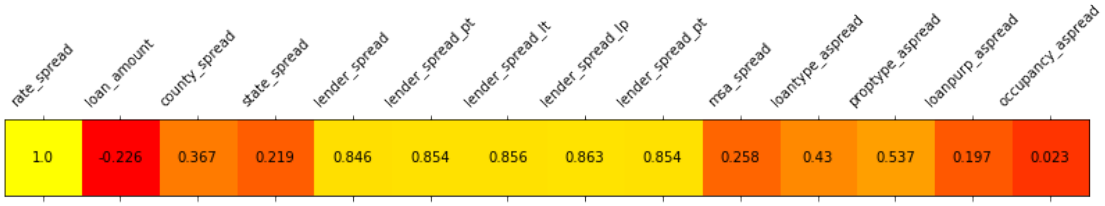
Below you can see an overview of the correlation between all major features. The more distant to zero (the higher the absolute value), the higher the correlation.



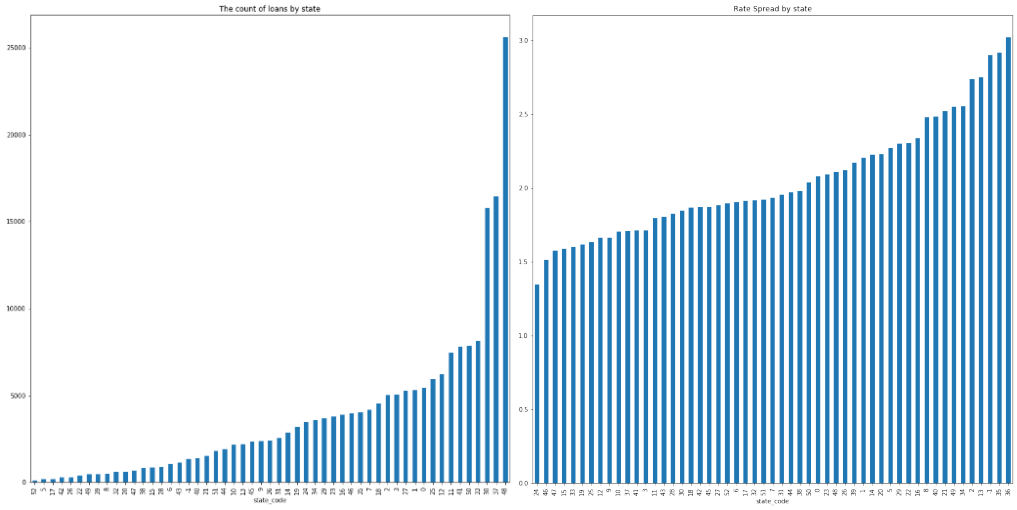
There is a small correlation between the amount lent and the rate spread and a slightly stronger to the income/loan relation but the strongest correlations are the previously aggregated averages by lending company, state and county. I further analyzed and got the following correlation heatmap to average rates by further groupings.



In the heatmap above there is basically negligible influence on rate_spread due to the applicant's race or ethnicity which could already be explained otherwise before with the preferences for the type of homes. There is a huge correlation though to the type of the home and the average rates of a specific lender and in a specific region:



Below you can see an overview of the huge difference of mortgage rates throughout the US. No state names were provided for the IDs but as you can see there is a quite uneven distribution with average rates up to 3% in certain states.



Cleaning process and removal of outliers

The exploration revealed that around 6.5% of the entries were incomplete and missing details about the applicant's income (5%) and the census information (1.5%) such as the average income in the metropolitan region. The essential features such as loan amount and rate were always provided. As the most often missing feature was the income, I trained a machine learning regression model with which tried to estimate it. Missing census features were enhanced with the county's mean.

Next to the missing values there were also some serious outliers in the data set, claiming mortgage rates of 99%, providing no income at all (zero) or incomes in the range of 20+ million dollars p.a. These values have been clamped if they exceeded 3 standard deviations of the feature's average.

Regression

The major goal of the project was to develop a machine learning algorithm which can predict the rate spread as accurate as possible. Overall, I engineered 68 different features, there off around 31 binary ones. These features were then scaled with the help of a fitted data normalizer to transform them into a -1 to +1 range and to ensure equal initial influence on the outcome by each feature.

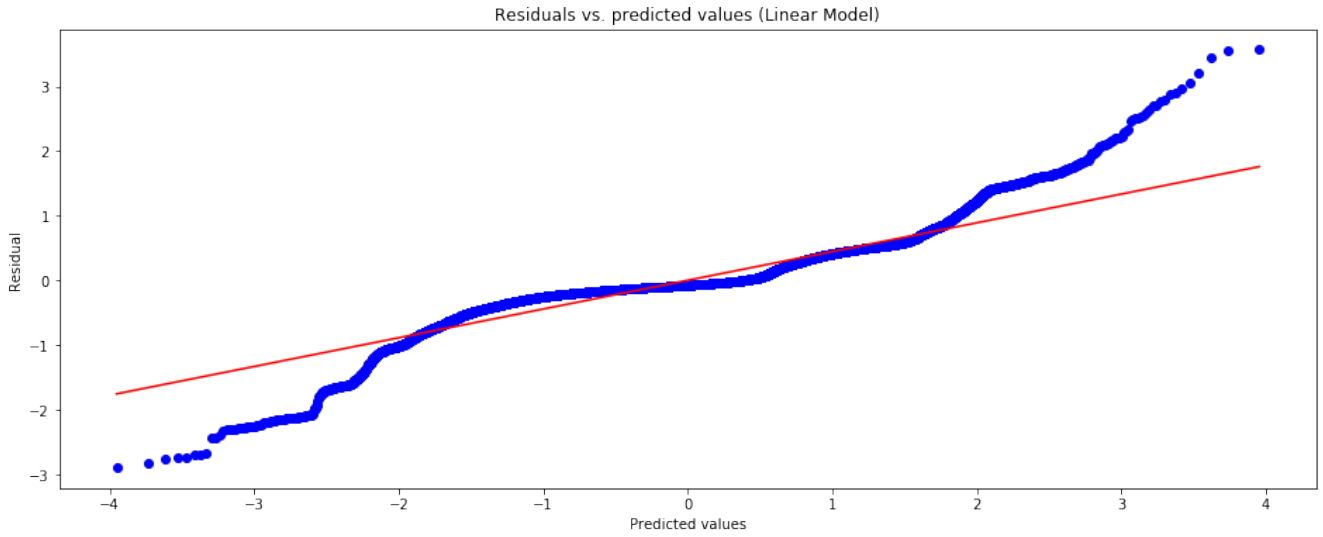
For the machine learning algorithm itself I evaluated three different approaches:

- A classic linear regression model (linear_model.LinearRegression)
- A neural network with (neural_network.MLPRegressor)
- A gradient boosting tree (ensemble.GradientBoostingRegressor)

Linear Regression

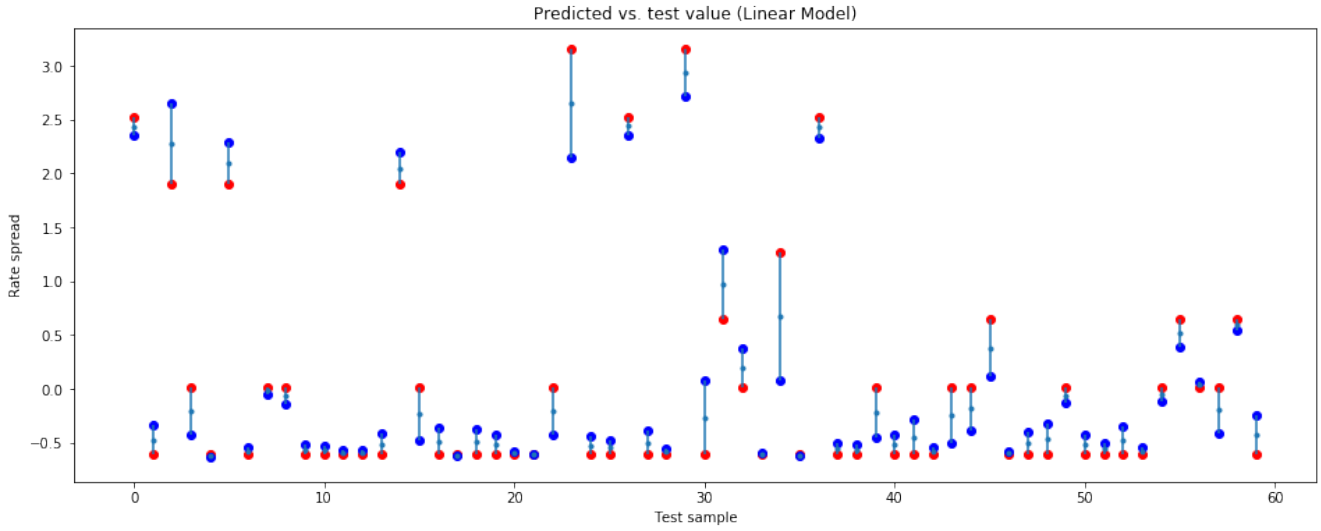
Due to the intensive data preparation and feature engineering even a simple linear regression model already proofed quite well with an R^2 score of 0.79 on the test set. Against the test set of the DrivenData competition contest site it though still performed effectively at around “only” around 0.74.

```
Mean Square Error      = 0.1975677717556343
Root Mean Square Error = 0.4444859635079991
Mean Absolute Error    = 0.2957563848737798
Median Absolute Error  = 0.17992695132983078
R^2                    = 0.7991129300603019
Adjusted R^2           = 0.7983623482129921
```

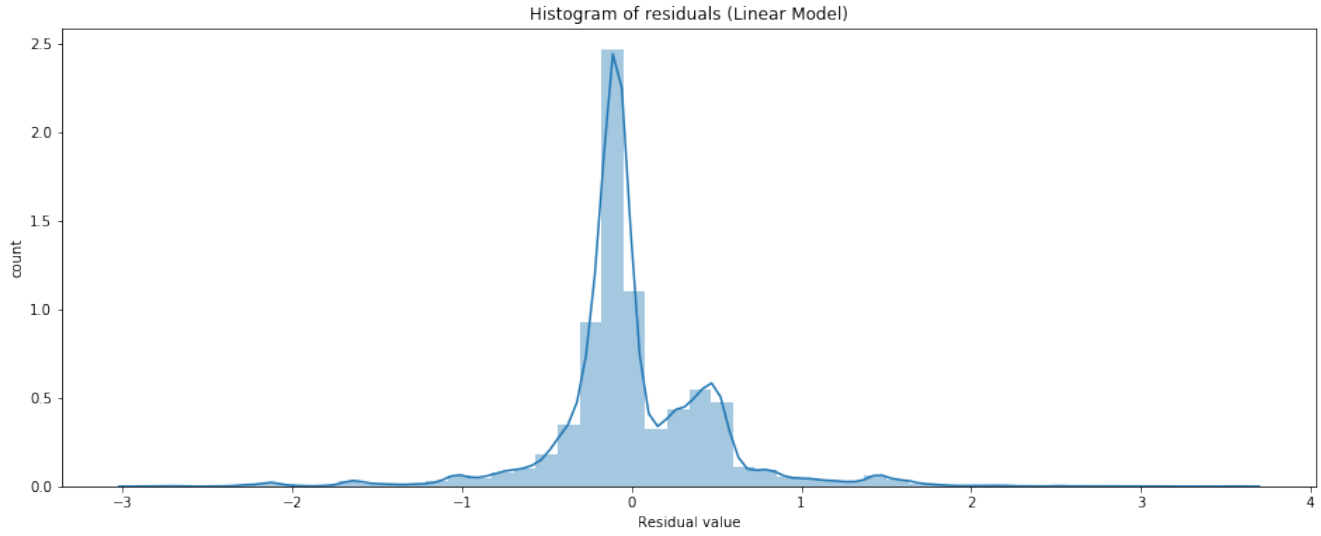


Majorly in the case of relatively high and low spread the model was off up to more than 4%.

Below you can see a sample of 60 predicted (blue) and test values (red). At least none of them is very far off, in the worst cases up to 1%, as factors such as property type and loan type at least provide a good baseline already.



A visualization of the residual error distribution. The average error rate is below 1%.

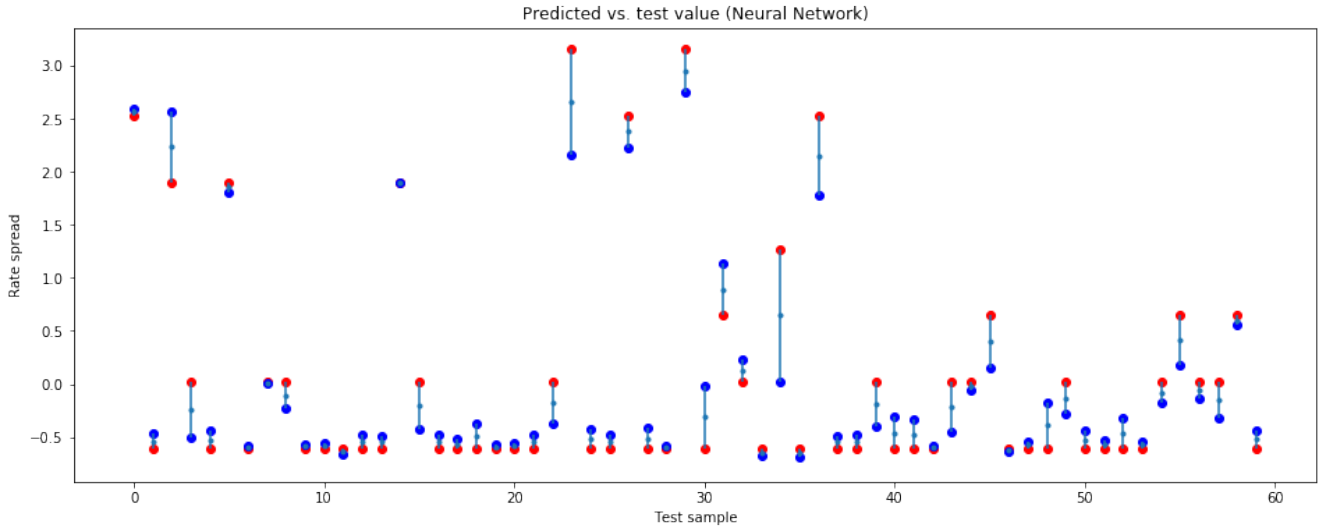


Neural Network

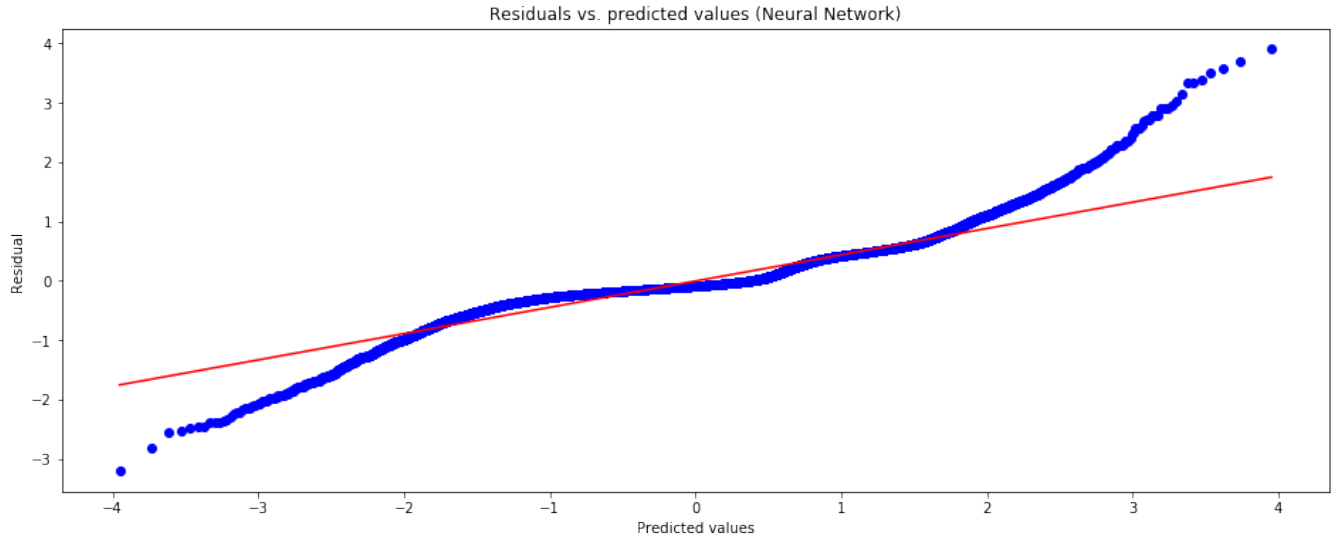
In the next step I built a neural network with 4 layers, 64 features as input layer, 2 hidden layers of size 32 and 8 and the mortgage rate spread as single output. I used the Adam optimizer and 500 training iterations with early stopping. This turned out to be the best set of hyperparameters. Increasing the count of layers above 2 resulted in overfitting, just one single layer did not let the model adapt to inter-feature relationships.
As just can see below an R^2 score of 0.801 performed slightly better than the previous linear approach. Before the intensive feature engineering the distance to the linear model was still far bigger though. This decreased the gap by an R^2 of around 0.07.

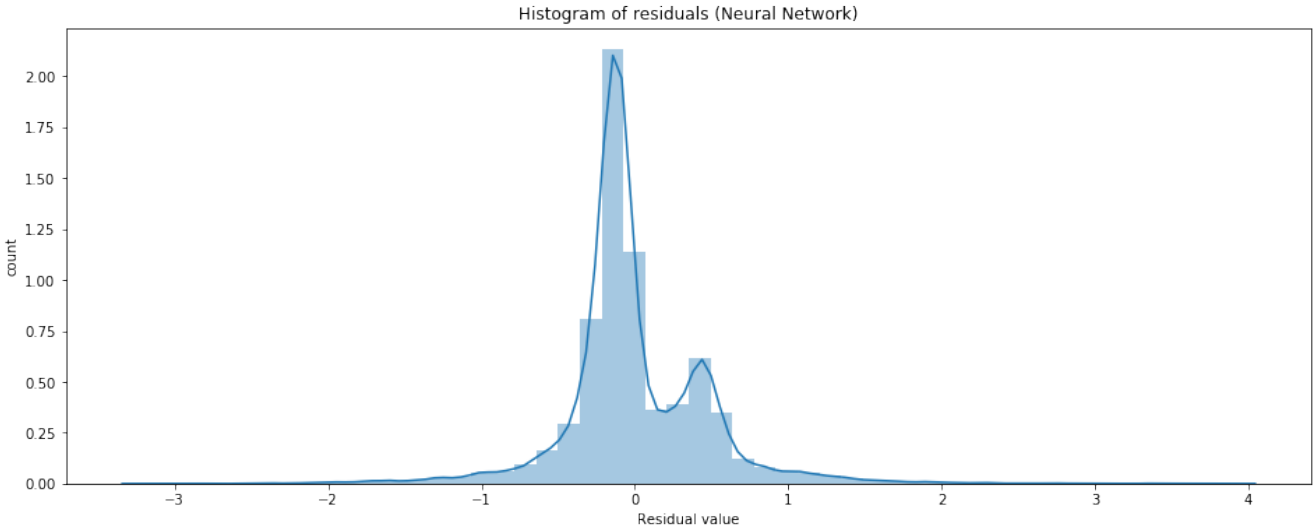
```
Mean Square Error      = 0.19481600592058723
Root Mean Square Error = 0.44137966187918903
Mean Absolute Error    = 0.29079334604780044
Median Absolute Error  = 0.17016910114178907
R^2                    = 0.8019109277845792
Adjusted R^2           = 0.8011708002004595
```

Below you can see the test and predicted relation is quite similar to the linear model



Here you can see the neural network's residual error curve. In comparison to the linear model it seems to have a lower variance and slightly higher bias, the curve is overall far smoother.

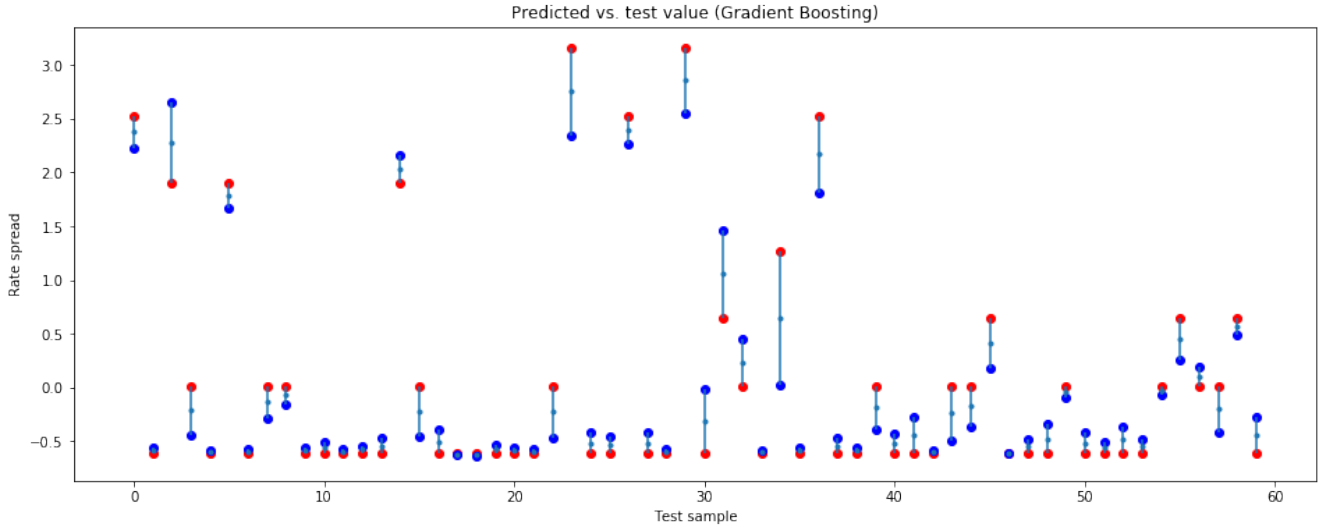


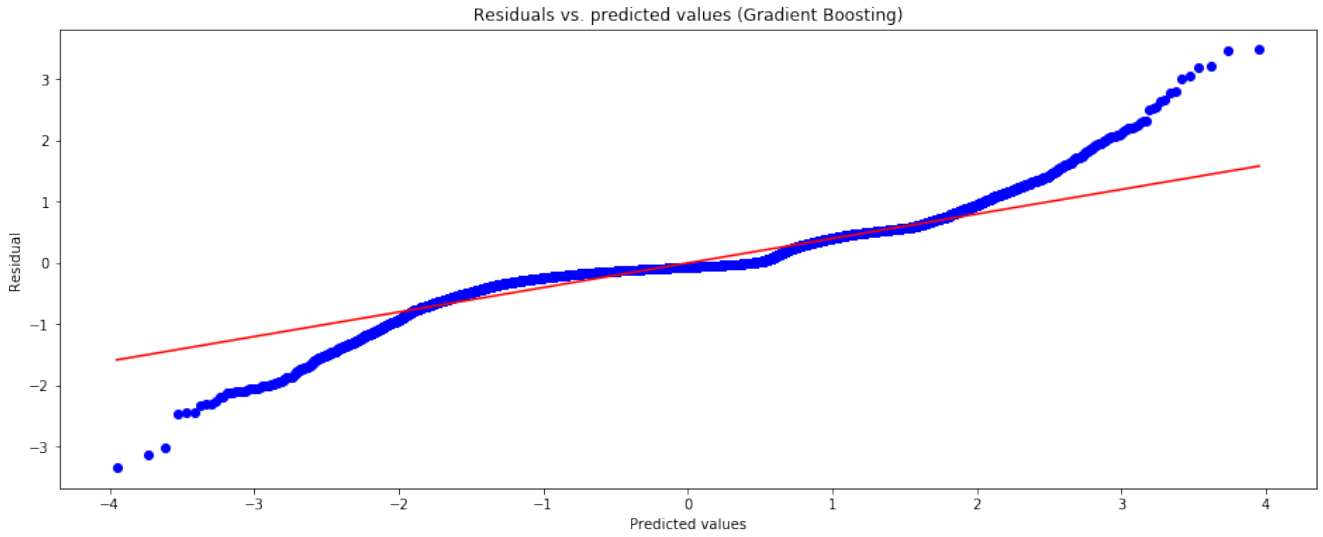


Gradient Boosting

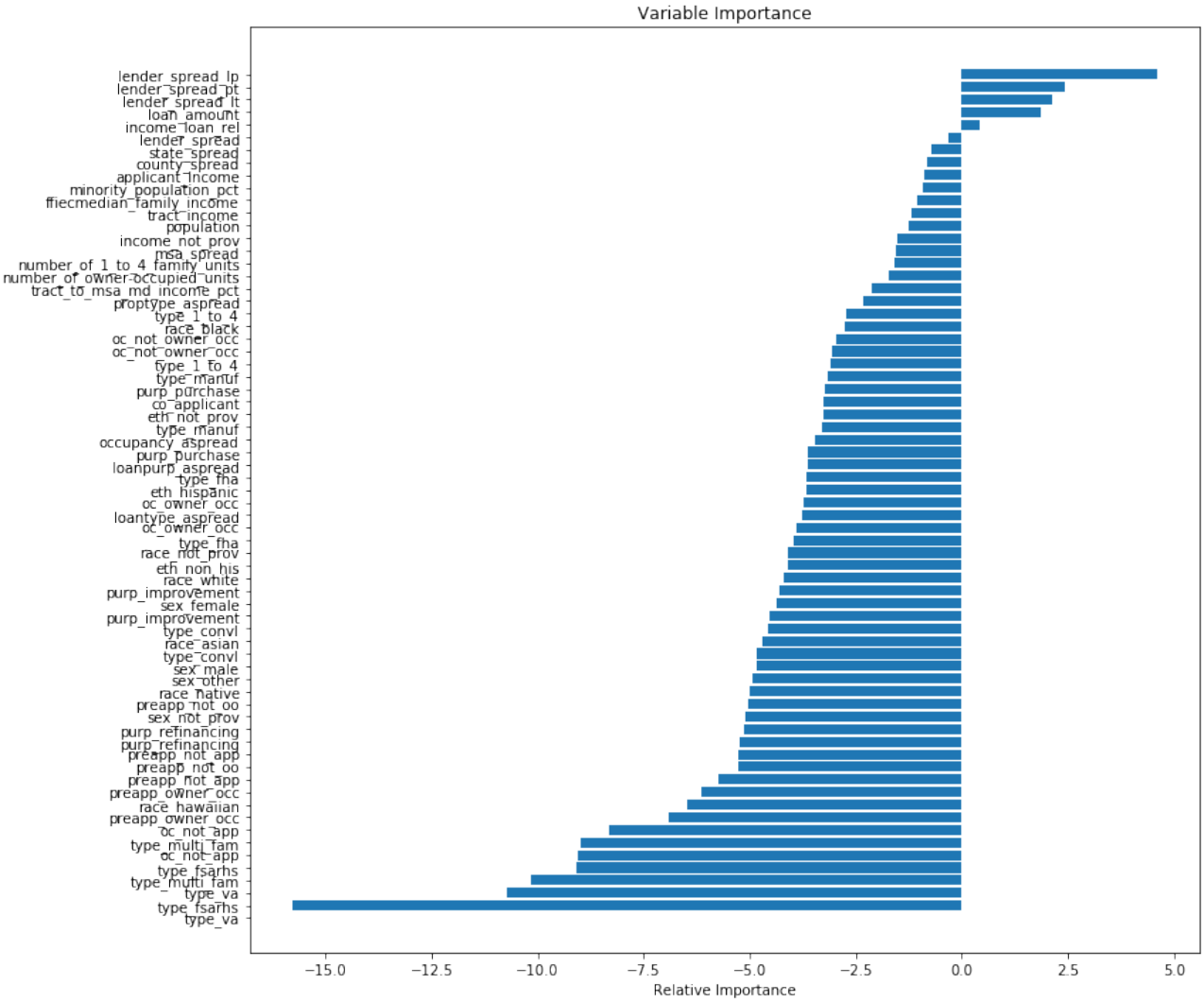
As third approach I tried an ensemble of gradient boosted trees. As much as deep learning models are praised - for good reason as shown above - boosting approaches are quite often the champions if it's about regression problems in Kaggle competitions.
One huge side-benefit they have over neural networks is that due to the fact that they are trees accumlating the final result they also let you analyze afterwards "why" they decided how they decided and which features influenced them how much in their decision.
In this project their reputation proofed true – theirs R^2 scored whole 0.02 points higher than the neural network did and instantly boosted my model - nomen est omen - into the top 20 in the DataDriven competition with a whopping R^2 score of 0.816 vs the internal test set and still real 0.765 vs Microsoft's competition set.
The only downside: Training the ensemble took easily an hour, further analyzing the weightings another two.

```
Mean Square Error      = 0.1798340104871867
Root Mean Square Error = 0.42406840307571453
Mean Absolute Error    = 0.2793412482001605
Median Absolute Error  = 0.16467731884914227
R^2                    = 0.8171446328454853
Adjusted R^2           = 0.8164614235214082
```





Below you can see on a logarithmic scale the hierarchy of the final tree ensemble:



Conclusion

The analysis proofed that **the loan application process in the US overall follows fair rules** and **does not show any measurable negative influence** on the mortgage rate offered **due to race, ethnical affinity or gender**. It showed that there is a significant variation of the rate in the US and between counties. In addition, it showed that white people and native Americans have a share of more than 15% for manufactured houses whereas this is no attractive choice for 96% of the Asian applicants.

The 5 most influential features affecting the mortgage rate provided are
1st. The lending institute's rates for specific loan types – the institutions seem to be specialized on specific loan and property types
2nd. The loan amount – larger loans usually get lower rates
3rd. The relation between income and loan amount
4th. The average rate within the state
5th. If the income was not provided at all the rate is significant higher – who does not grant permission to insights into his/her financial situation might have something to hide.

It turned out to be very effective to provide the machine learning algorithm lots of features aggregated from the means of groupings by category such as loan types, house types, counties, states and lenders. Especially the further partition of customers of a single lender into more detailed groups brought a huge improvement. This enabled the machine learning models to focus on the right balance of the influences of this single, categorical groups onto the final output instead of by itself having to find complex relationships in this huge data set.

Overall the boosting model proofed to be the best performing solution. In the DrivenData.com competition it reached place 6 of 709 with a score of 0.7703:

BEST SCORE	CURRENT RANK	# COMPETITORS	SUBS. TODAY
0.7703	6	709	1 / 3