DETECTING BIAS AGAINST MINORITIES IN HOME LOAN APPLICATIONS

A PROJECT REPORT by **LOAN SHARKS**

Contributors

- **Anantanarayanan G Iyengar** A20388360
- **Omkar Pawar** A20448802
- **Virat Joshi** A20417850
- **Xiaoman Shen** A20449626
- **Bhuvnesh Tejwani** A20444878

ABSTRACT

The goal of this project is to analyze if there are discriminatory policies followed in home loan lending by banks and other financial institutions. We plan to analyze conventional home loans for this project. A conventional home loan is a loan that is not backed by FHA/VA, etc. We will build models for different states which include African American and minority neighborhoods and white neighborhoods using data provided by the Consumer Finance Protection Bureau. These models will allow us to predict the probability of an individual obtaining a conventional loan in a particular minority neighborhood vs a similar loan application in white and other wealthier neighborhoods. This comparison will help us realize if there is an inherent bias in the loan approval process against specific communities along with the reason for such a bias.

BACKGROUND

There have been a number of reports like the following:

- NPR Gap between white and black homeownership in Baltimore
- UrbanWire. What explains the gap between black and white young adults
- CNN. Racial bias cost
- Why does a homeownership gap exist between whites and minorities

The overarching theme remains the same. Minorities such as African Americans, Hispanics, Latinos, Asians, etc continue to struggle to obtain financing for purchasing/refinancing homes at a higher rate when compared with whites.

Homeownership lies at the heart of the American dream. In the US, wealth and financial stability are linked with homeownership. This is for good reasons. As per this article, homeownership is still financially better than renting. In general, owning a home for a reasonable period typically more than 10 years has historically outperformed the stock market and provided better returns for homeowners. As per the survey of consumer finances, the average homeowner has a net worth of 231K while the average renter has a net worth of 5K. Homeownership helps build communities by boosting investments in housing, construction, which in turn helps drive local economies. This, in turn, helps intangibles like reducing crime, etc.

In the US and Canada redlining is the term which basically means systematic denial of various services like loans and other forms of financing to specific communities on the basis of race for the most part. The term redlining for financing means marking off certain zip codes/areas, etc as high risk. Banks and other financial lending institutions would make it extremely difficult for home loan borrowers to get loans for homes in these neighborhoods.

The Federal fair housing act and the Community Reinvestment Act were laws passed by Congress, the latter in 1977 to encourage banks and other institutions to reduce discriminatory practices against low income and minority neighborhoods. While the intent behind these laws should be applauded, laws need oversight and strict enforcement to catch perpetrators who violate them. The laws need to be updated for the modern era where loan decisions are made by machine learning models and AI.

DATA SOURCES

Our primary data source is the HMDA data provided by the Consumer Protection Financial Bureau for the years 2007-08 and 2014-17. The early parts of 2007 were when the housing bubble fueled by the subprime mortgage boom was at its peak. This gives us some insights into whether there was a difference in lending patterns in the 2007-2008 period vs recent periods.
Link to the dataset - https://www.consumerfinance.gov/data-research/hmda/historic-data/

We can download the data statewide and year wise depending on the requirement. As mentioned previously, we use statewise data for the years 2007-08 and 2014-17 for the states of Pennsylvania and Illinois. These states were chosen as they have a sizable minority population.

The data here describes if the consumer got the mortgage—look for applications that were "originated"—or if the consumer was denied, didn't complete the application, or something else happened. For each record, it describes the loan, the property characteristics, the applicant demographics, and the lender. The datasets consist of 310000 rows on an average and 78 columns.

We have also considered two additional data sources such as the Zillow Home Value dataset and the United States Census Bureau.

The Zillow Home Value dataset is provided by Zillow and provides us with the monthly median home values by region and type. It contains information such as region id, state, and municipal FIPS codes and monthly median prices. Banks use the home value to loan ratio as one of the predictors in their decision making process. HMDA data does not have this information. We felt that using the median home values in a county for this purpose would provide us with a reasonable approximation.

We used the United States Census Bureau dataset which provides us race and demographic information per county. While the HMDA dataset does include the minority population in a county, we wanted to see if the percentage of specific communities like African Americans, Asians, Hispanics, etc in a county/tract was a significant factor in the outcome of a loan application. Apart from the state and county codes, this data also contains information such as county name, the total population, and gender distribution for different races and ethnicities.

We have grouped all the datasets used throughout the project. It includes raw data and imputed datasets as well. It can be found here.

Zillow Home Value Dataset - https://www.zillow.com/research/data/
Census dataset - https://www.census.gov/data/datasets/2010/dec/summary-file-1.html

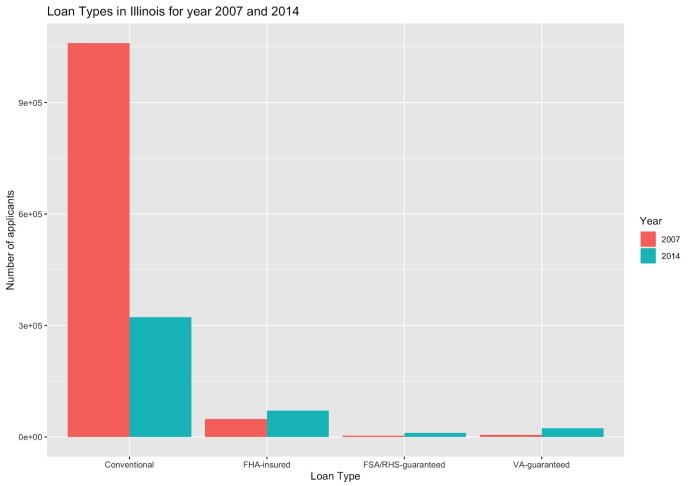
EXPLORATORY DATA ANALYSIS

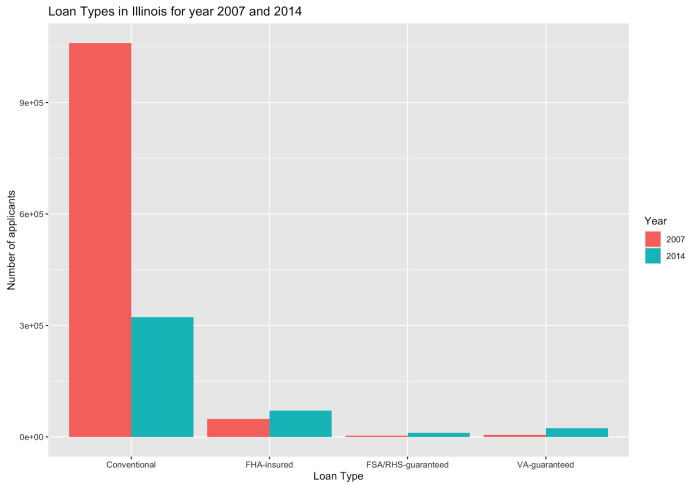
We want the data to tell us the story. To check if there is an implicit bias in the home loan lending models we need to explore the dataset and periodically check if we are moving in the right direction. Digging deeper will help us do the same and we start by checking some facts about the HMDA dataset.

As we know, the dataset is from two states namely Pennsylvania and Illinois and has 78 columns in total for both states. The number of observations vary state and year wise in the range of 300000 to 600000. We perform an analysis of these observations based on specific parameters starting with Loan Types.

- **Loan Types**

We look at the number of applications made in the years 2007 and 2014. 2007 is an important year to consider in our analysis since it was the peak of the housing recession in the United states and it provides an insight into the buying habits of Americans in different scenarios. The graph below shows that the number of home loan applications remain high even in 2007 and conventional loans are the preferred type. So, we will draw our attention to conventional loans for further analysis.

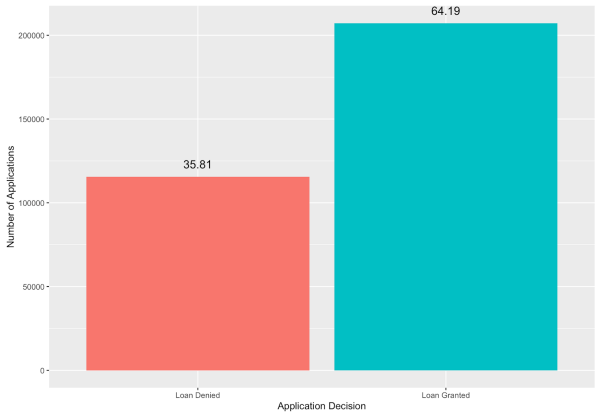




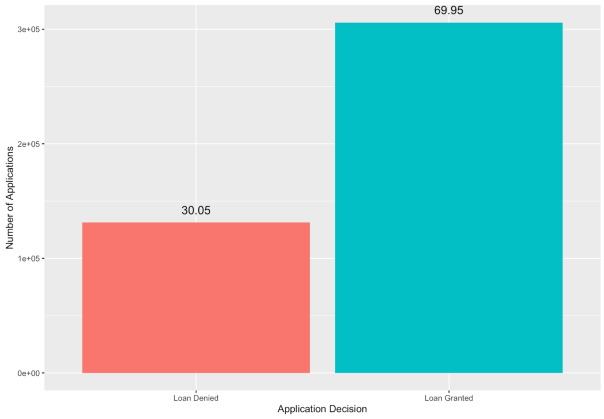
- **Proportion of Loans Granted and Denied**

For conventional loans, we need to check the ratio in which they were denied and granted in both the states. Illinois had a higher loan approval rate of 69% as compared to Pennsylvania where the approval rate was 65%.

Pennsylvania

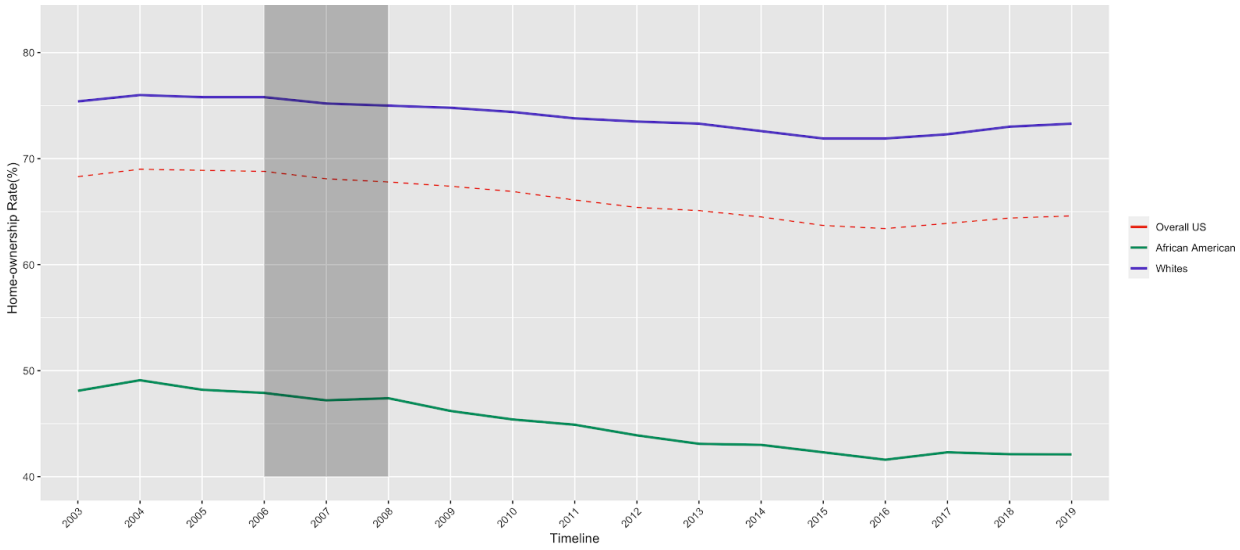


Illinois



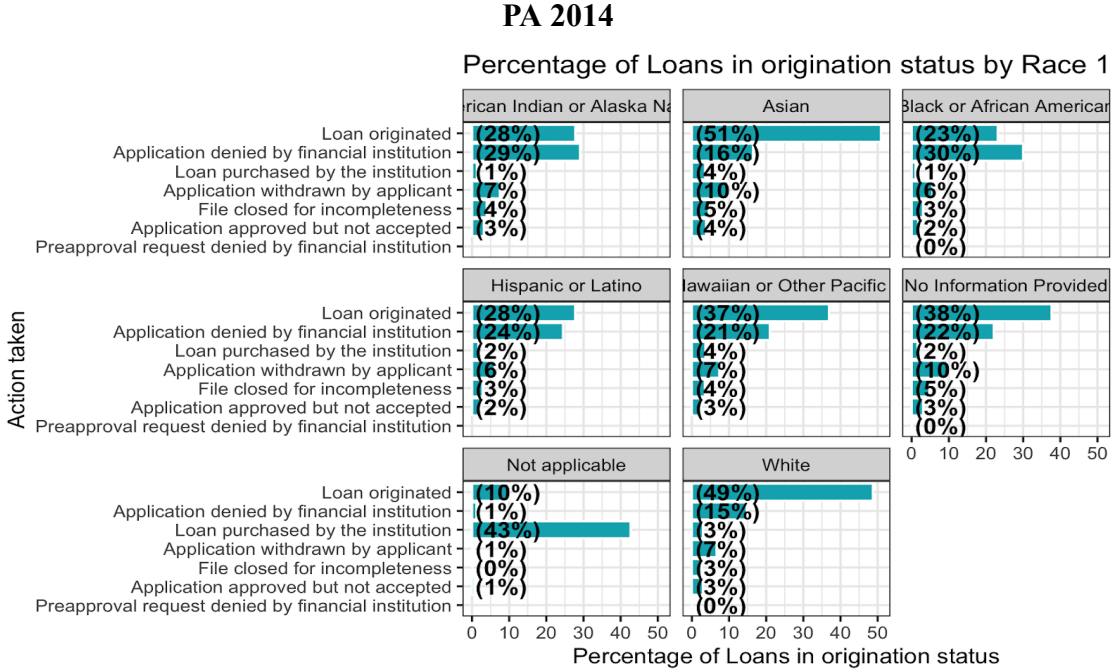
- **Homeownership Share**

We look at the home ownership share for the years 2003 through 2019 and it can be seen that the minorities have a significantly lower share as compared to the total US population. This share dropped further in the years 2007-08 indicating the effects of recession. It is also evident that the white population has a higher share in contrast to African Americans. The most interesting insight is that the share for white population has a minimal effect during the recession years.



- **Action Type According to Race**

A categorized comparison of the application decisions based on the parameters of race and ethnicity tells us that the denial rates are higher for African Americans and other minorities. Given that 2007 was the peak of the housing boom, we feel that it is possible that a high proportion of mortgage applications for minorities might have missed the race information. The plot below shows a high number of applications where race information wasn't provided.



DATA CLEANING

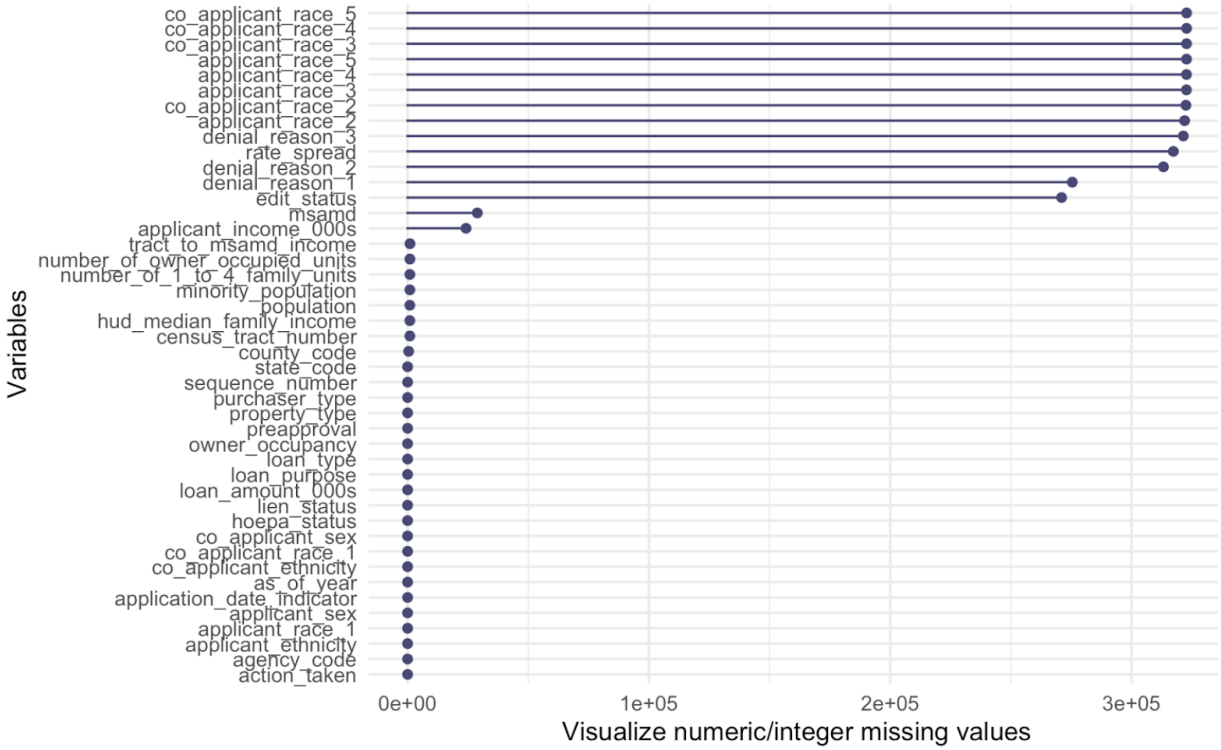
The data we use has 78 columns but not all of them are necessary to our analysis. As a result we need to perform data cleaning. The approaches used by us for the same are listed below.

- **Removing Redundancy**

Most columns in the dataset provide the same information in different formats. For example, we had county_name and county_code columns. Both these columns provide the same information, just one column represents it in the form of codes and the other in the form of strings. These redundancies were removed by selecting only the desired columns and dropping the rest in the 'utils' file. This reduced the dimensionality of the data and made it easier to extract useful information from the remaining columns.

- **Checking for Missing Values - Column Wise**

Being a raw dataset, HMDA data had a lot of missing values in various columns. To find the missing values, we looked for NAs, NULL values, empty fields or any values that were coded "?" as missing. There were no missing values as NULL or "?". But we found values that were coded as NAs and a lot of empty strings. Following is the missing values plot for the dataset



It can be noticed that the race values for much of the columns are missing. These columns convey additional information about race. We felt that the primary race column, which is applicant_race_1 and co_applicant_race_1 were enough for our analysis. Hence these columns were dropped from our analysis. We can see from the above plots for missing values for Whites vs African Americans, that the pattern of missing values is generally random.

- **Checking for Missing Values - Row Wise**

We also tried finding missing values from each row. In this approach, we found the percentage of missing values in each row. Say there are 100 columns and 90 columns for a single observation are empty. It says that 90% of the information is missing for that observation. Thinking about that, this particular observation did not provide adequate information as a whole, so it can be disregarded. There were around 5% of such rows in our dataset. Getting rid of these rows helped us bring the number of NAs per column down significantly. Other than the applicant's race, applicant's income is also missing and we need to fix this accordingly. This is an important predictor for us, hence we carefully review this column, its distribution and fill missing values as needed.

First we used **mean** to fill the missing values, but it turned out that the distribution is highly skewed towards the right. Hence, mean was pulled more towards the right. It was rational to fill these values with **median**. So we tried that too. But then when we tried modelling our data with these imputations, the performance was not as expected.

Hence, the next option that we tried to fill in the missing values is by building a model and predicting the missing values with ***mice*** in R. Imputations using mice were quite time consuming

but proved to be better than any other approaches. Hence, we used mice for most part of the missing data imputations.

To summarise the data cleaning process, we took the following steps

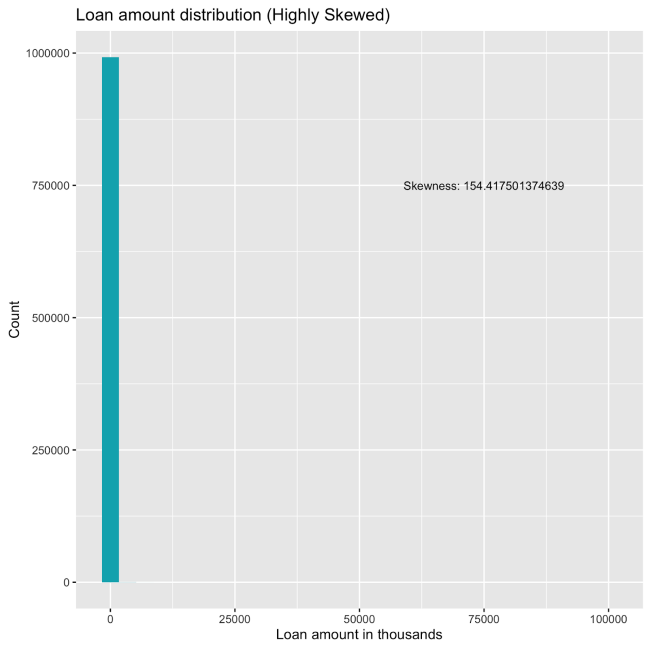
- Removed redundancy
- Found Missing Values encoded in different formats
- Row Wise Missing values count
- Replace by Mean (Did not work well for highly skewed predictors)
- Replace by Median (Did better but distribution peaked at the median)
- Impute using Mice in R (Better approach than any other practices incorporated before.)

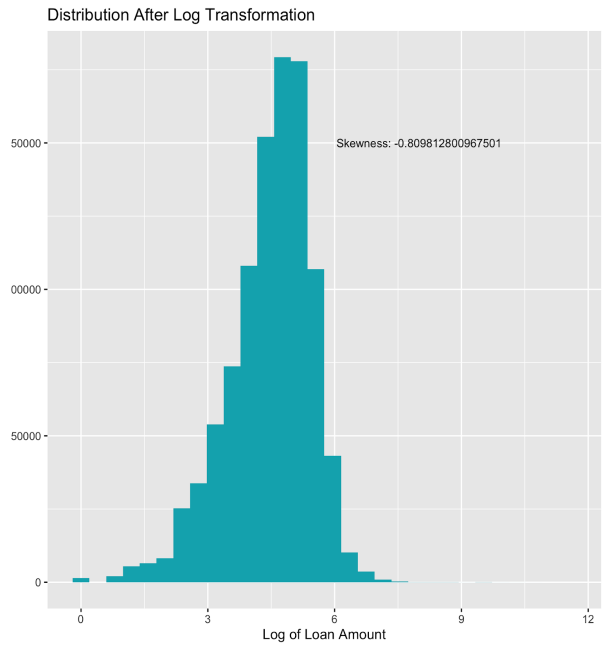
FEATURE ENGINEERING

This phase of the project involves improvising the features that we have in our data and adding new features on the basis of existing ones using domain knowledge which can facilitate the improvement in model performance. We apply few transformation operations on the predictors so as to get them in required distribution. Let's walk through the steps that we took in feature engineering.

- **Feature Scaling**

A couple of numeric variables like applicant income and loan amount were highly skewed. The reason that we see this skewness is due to the fact that a majority of people have income in a certain interval, but there are cases where people have income even more that $1000K. So it is pretty obvious to see this skewness in such predictors. We were planning to use modelling techniques which assumes that the data is normal. Hence, to make it normal, we used **log-transformation.** Scaling these predictor variables is always useful and we get better performance. The following plots for a loan_amount show how log transformation turns a skewed distribution into a normal one.





Same can be observed with the applicant's income distribution.

- **Creating New Features**

All the features that we need to make the decision of any application of loan were not present in HMDA data. Before we merge other datasets into this, we thought of creating some new features that could give us some extra information or reduce the complexity of that predictor.

- **Loan to Income Ratio**

Loan to income ratio, as the name suggests, is the ratio of the loan amount and the income of that applicant. While we researched the loan granting process, we came across this metric where banks calculate this ratio to see the feasibility of the application, meaning how much money the applicant is asking the bank to lend and how much he can really pay back. This predictor was missing in our data and hence we decided to add it in as a new feature.

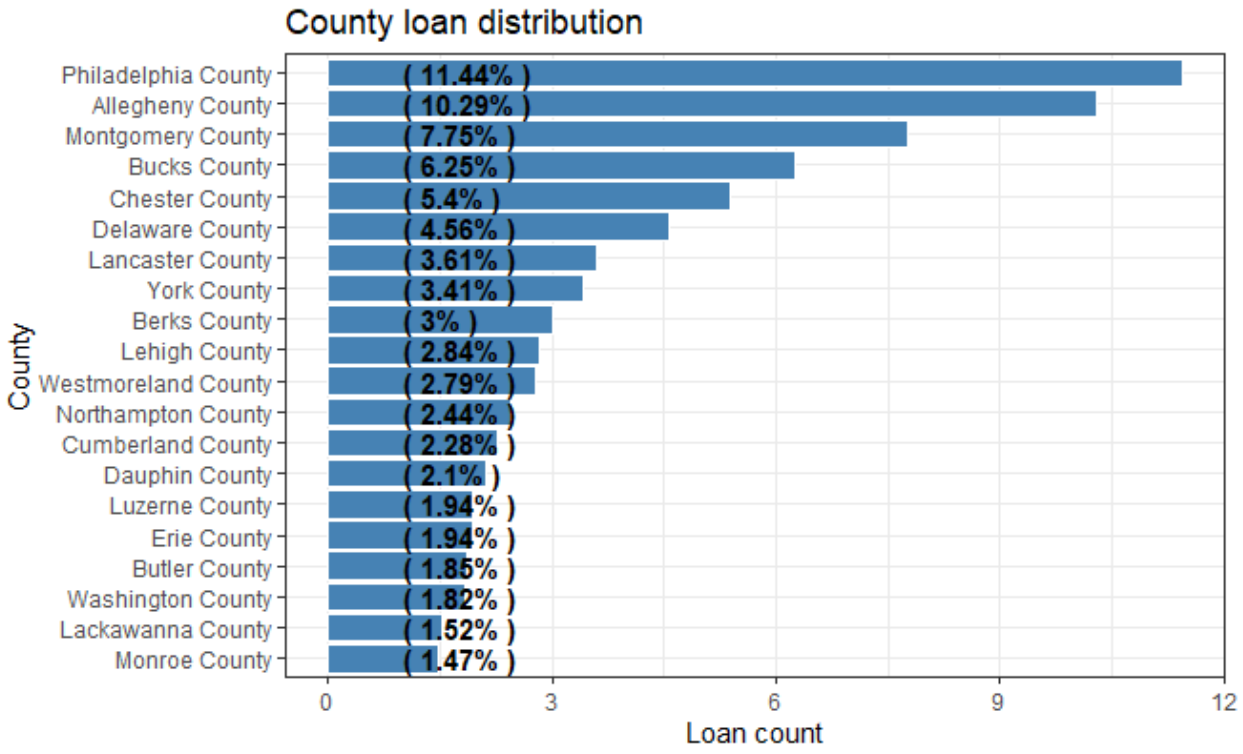
- **Applicant Race and Ethnicity**

HMDA data has two columns specifying race and ethnicity of the applicant. US Census defines race as a person's identification with one or more social groups. For e.g. White, Black or African American, Hispanic, etc. Ethnicity determines whether a person is of Hispanic origin or not.
An applicant can report their race as White and ethnicity as Hispanic. They would not be categorized as White in the loan applications. Hence we merged the race and ethnicity of the applicants into one column.

The hierarchy is that we first split the applicant into “Hispanic or Latino” according to the ethnicity. If the applicant does not belong to this category , this new column gets the race name as its value.

- **Capping County Names/County ids.**

County names are treated as factor variables. In a state like Pennslvyania, there are around 60+ counties, more than that in Illinois. If we build a model using a factor variable with values this large, it would take forever to run. Hence, we decided to cap the county names. We found the top 25 counties from where most applications came in both the states. Counties other than top 25 were grouped into “Other”. Following plot shows how it makes sense to cap the top few counties and group the rest of them in others.



These are top 25 counties in PA from where most of the loan applications came. The number on each bar shows what percent of the applications were made from that particular county. Almost **75%** of the applications came from these counties.

- **Loan Granted**

We created a new column which holds the result of the loan application process. This column was set to true based on the value of the Action type field. For simplicity we followed the following logic.

We filtered the dataset to exclude the following action types.

1. Application withdrawn by applicant
2. Loan purchased by the institution
3. Application approved but not accepted.

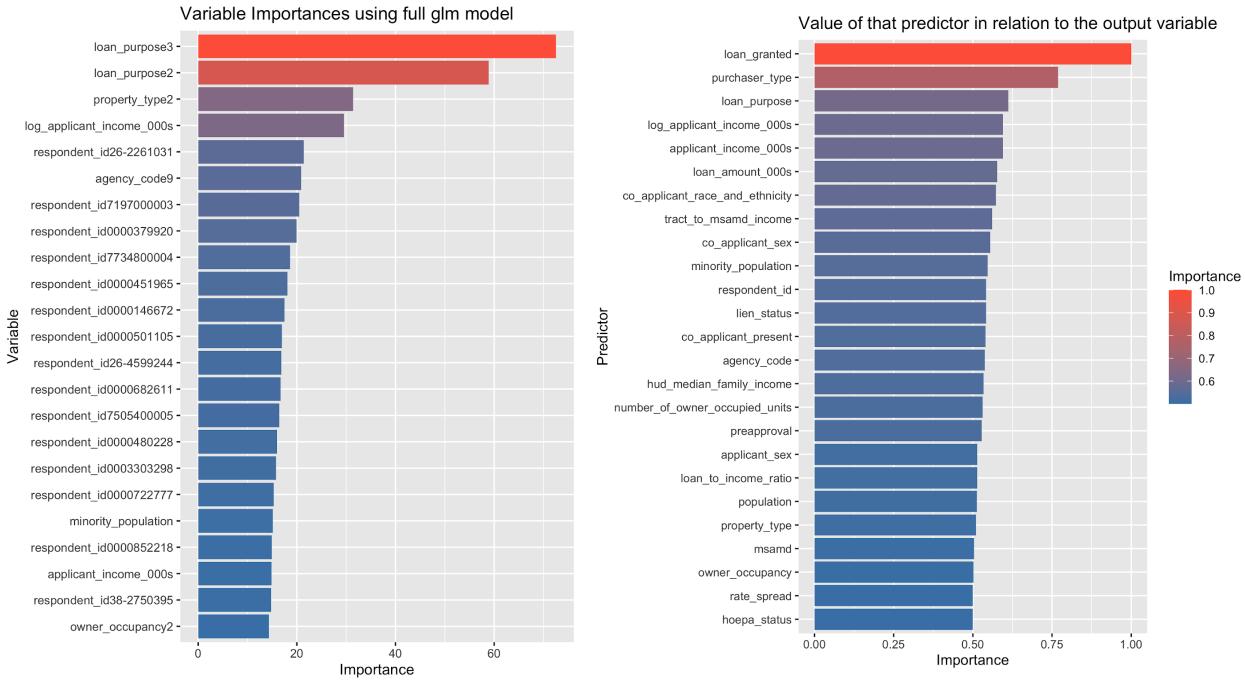
We retained the following action types.

1. Loan originated
2. Application denied by the financial institution.
3. File closed for incompleteness.
4. Pre Approval request denied by financial institution.

Please note that it is possible that banks are known to take longer to decide on loan applications for minorities. This could result in the applicants withdrawing their applications. For simplicity we decided to ignore these action types.

- **Feature Selection**

Variable Importances according to glm



We did not select the purchaser type predictor column in our models as that conveys information about a loan being approved.

DATA MODELLING

After we had done enough data cleaning, we started modeling our data. Being a supervised classification problem, we tried 4 different approaches to model our data which suits these kind of problems

- Naive Bayes
- Decision Tree
- Random Forest
- Logistic Regression.

Each model is different and we selected only these models for a reason. The reasons for the same are mentioned while we discuss each model briefly. But before we get into it, let's discuss what performance metric we used to gauge the performance of these models.

- **Performance Metrics**

To gauge the performance of our models, we used evaluation metrics that are mentioned below

- **Precision**: Precision is the ratio of correctly predicted positive observations to the total predicted positive observations.
- **Recall**: Recall is the ratio of correctly predicted positive observations to the all observations in actual class - yes.
- **F1 Score**: F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account.

Each of them are interdependent and we need to decide which value to maximize. Our data is imbalanced, so just accuracy may not tell us the full story. For our problem, we choose to maximize precision keeping decent recall values. Positive class for our problem is loan granted. So, thinking from the bank's perspective, maximizing precision is equivalent to minimize the risk that the bank takes to make a decision about an application. There is always a tradeoff between precision and recall. We can gauge this by changing the threshold values.

We have selected the following predictors throughout the modelling process. Description for each predictor is given.

Predictors

Name	Type	Details
applicant_race_and_ethnicity	Factor	Details about the race and ethnicity of the applicant. For e.g. White, Black or African American, Asian, etc.

owner_occupancy	Factor	Details about the intent behind the purchase, i.e. whether the applicant intends to stay in the property, rent it out, etc.
preapproval	Factor	Details about whether the loan was preapproved, etc
log_loan_amount_000s	Numeric	Log of the loan amount rounded to the nearest thousands.
applicant_sex	Factor	Details about the applicant sex
county_code	Factor	Identifies the county. Please note that we capped this to the top 25 counties in the state. For the rest please enter the value as Other
tract_to_msamd_income	Numeric	Ratio of the census tract income to the metropolitan division income.
co_applicant_present	Factor	Boolean value indicating whether there a co applicant was part of the application
agency_code	Factor	Identifies the regulatory agency
minority_population	Numeric	Proportion of the minority population in the tract
loan_purpose	Factor	Purpose of the loan, purchase, refinance, home improvement, etc
rate_spread	Numeric	Deviation from the prevailing interest rate
property_type	Factor	Type of the property, single family home, town home, etc.
loan_to_income_ratio	Numeric	Ratio of the loan amount to applicant income
respondent_id	Factor	Identifies the financial institution. Please note that this was capped to include the

		top 50 institutions in the state.

Now that we have discussed what predictors we will be using and what are our performance metrics, let's see the results for each approach.

***Note**: We have implemented models on data from PA and IL. While we discuss the performance of a model, we consider only one specific year and one state. If you wish to see models which are not mentioned below, please refer to the repository. You can find it in src>train and specific year.*

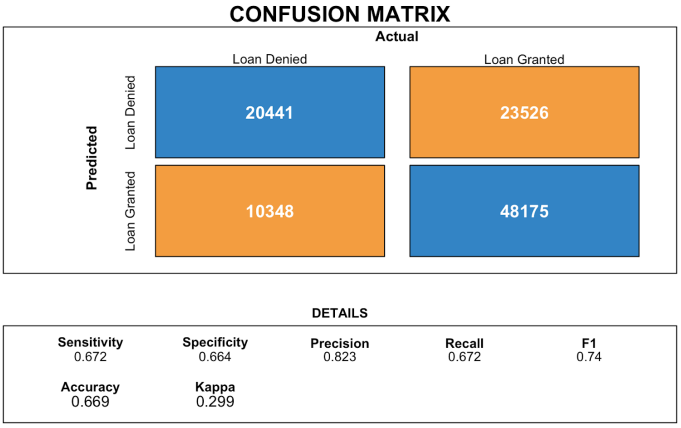
- ## Naive Bayes

It uses Bayes theorem, which is based on conditional probability concepts to classify the observation into a specific class.

The performance of this model can be used as a benchmark to compare how other complex models perform. To set a benchmark, we built a Naive Bayes model on data from PA 2014 and 2015 merged.

Model Performance

Looking at the confusion matrix, we see that Naive Bayes does not perform better. We can see that the precision value is good, but Recall in not acceptable, very low. This is the simplest model of all and we have set the benchmark. We move ahead and start implementing some complex algorithms and see how they perform.



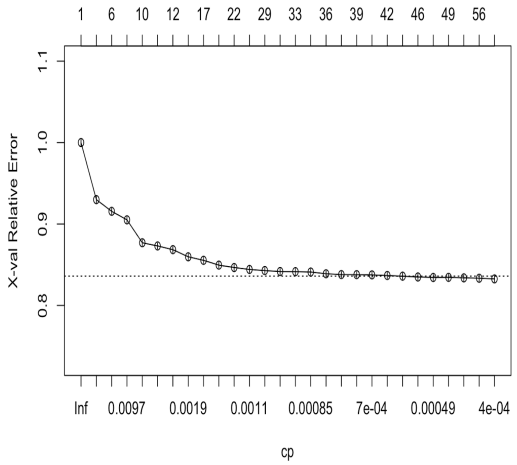
- ## Decision Tree

It is a non-parametric model which trains on observed data and needs minimum data transformation. In addition to that, decision trees closely mimic the human decision-making process. So, it is easy to understand the flow of this algorithm and know what features from your data drive the decision-making process.

Hyperparamters:

We tuned the minimum split, cp, the complexity parameter, and max depth. The values for the optimal tree for these parameters are shown below.

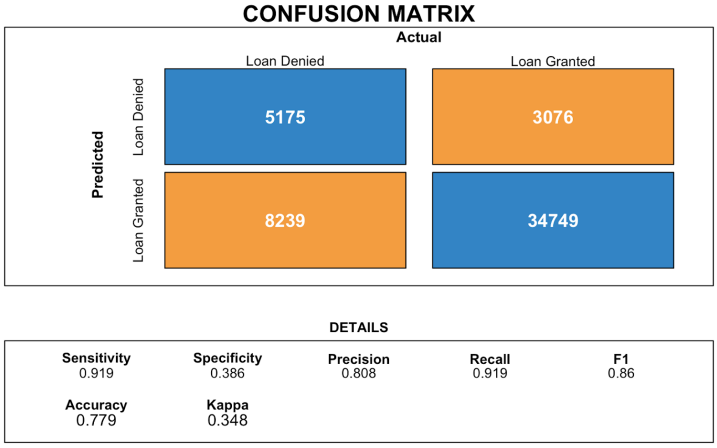
Hyperparameters	Range Tested	Optimal Value	Description
Minimum Split	2-50	14	Minimum number of observation to further split a node
CP	0-0.05	0.0007	Complexity Parameter. Restricts the growth of tree
Maximum Depth	10-40	20	Max depth controls how much tree can grow deeper.



The graph alongside shows that if we keep reducing the value of cp any further than 0.0008, the model does not perform any better. Hence it is not worth adding more complexity to the tree

Model Performance

The confusion matrix alongside is for PA 2014. We can see that the model performs better than Naive Bayes. Also, we notice that Precision and Recall is high for the same.



Taking that into consideration, we can look at the tree and see if Race and Ethnicity is having any effect on the decision making for loan application.When we plot the decision tree, it can be observed that Race and Ethnicity has significant role in decision making and one of the key feature that drives the decision. The decision tree diagram is added in the appendix for reference.

So, using this approach, we found evidence that models do take into consideration race and ethnicity values. When we try to remove this feature from our model, the performance degrades, showing that what we are trying to investigate is true.

This was all about decision trees. Let's move forward with our next approach.

- ## Random Forest

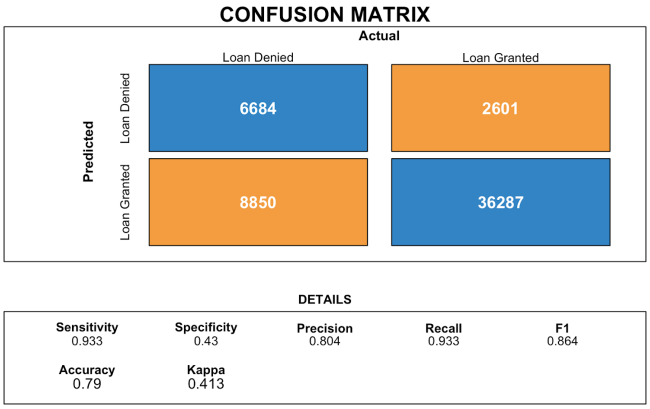
This algorithm is built on top of decision trees. It is a collection of many decision trees where each tree gives an output and the results of all these trees are combined to get the final output. It is more complex than a decision tree, but we can try to see if it does any better. We don't discuss much about this approach here, but it gives us an idea about how adding more trees affects the performance of the model.

Random Forests take longer to train as it is more complex. We try to avoid higher training time if the performance does not show significant improvement. Lets see how it performs.

Model Performance

Looking at the results of Random Forest from PA 2015, we can see that it shows almost the same performance as that of Decision Trees. Hence it is not worth adding that amount of complexity and get very little improvement of performance.

Hence, we move forward with our next and final model.



- ## Logistic Regression

We chose this classic model for a reason that we can interpret each variable independently by keeping others constant. In this way we can find whether race and ethnicity is having any effect on the results that we predict. Also, this is the final model that we deployed. We get probability for each class for every observation. Taking these probabilities into account, we can vary the threshold to tune the precision and recall parameters.

One unique benefit of using these kinds of approaches is that we can keep varying the coefficient values for a single predictor by keeping others constant. This way, we can find how much that particular predictor affects the decision.

Also, logistic regression facilitates the concept of the odds ratio. When we build a logistic regression model and find the coefficients for each predictor, using these coefficient values, we

can find the odds of getting the loan approved for a particular category . It is discussed below in the results interpretation section.

Logistic regression performs better when the continuous predictors have a normal distribution. Following is the list of our best models for logistic regression. The best performing models are the ones that take log transformed values for applicant income and loan amount.

- Log transformed loan amount and applicant income
- Log transformed loan amount and loan to income ratio
- Log transformed loan amount and loan to income ratio with census race data.
- Log transformed loan amount and loan to income ratio with median loan to home value ratio from Zillow.

We selected model #2 from the list above (Log transformed loan amount and loan to income ratio). The models with census race data and the median home to loan value ratio did show promise. However we felt that the gains were limited and hence we went with the simpler model. For details on these models and their performance please refer to the sections after the appendix.

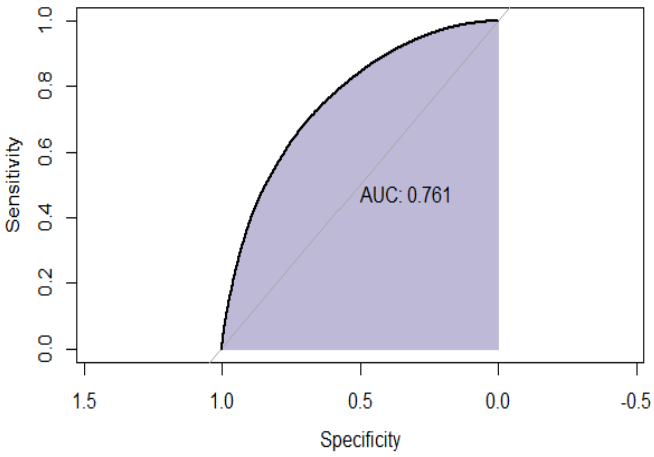
The model details for Pennsylvania 2014 are given below. We use ANOVA on a binomial model with a logit link and present it as an analysis of the deviance table. The response is 'loan_granted' and the terms are added sequentially from first to last

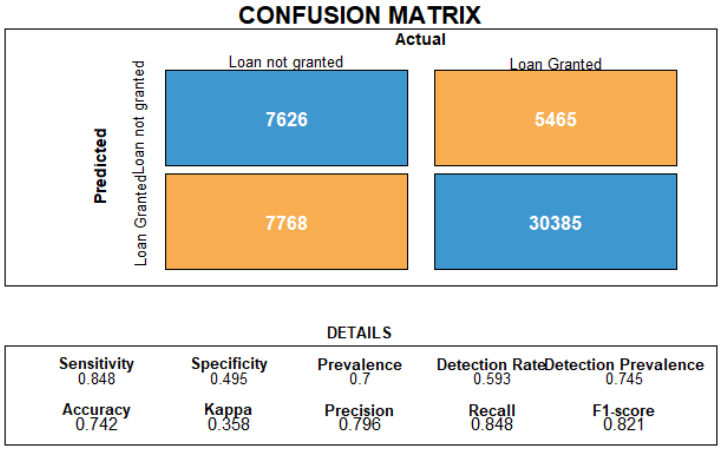
	Df	Deviance	Resid. Df	Resid. Dev
NULL			204982	250576
applicant_race_and_ethnicity	7	7939.9	204975	242636
owner_occupancy	2	35.4	204973	242601
preapproval	2	3977.1	204971	238624
log_loan_amount_000s	1	3635.9	204970	234988
applicant_sex	3	185.6	204967	234802
county_code	25	1004.5	204942	233797
tract_to_msamd_income	1	768.7	204941	233029
co_applicant_present	1	883.0	204940	232146
agency_code	5	3585.3	204935	228560
minority_population	1	107.3	204934	228453
loan_purpose	2	6343.0	204932	222110
rate_spread	1	418.8	204931	221691
property_type	2	1152.7	204929	220539
loan_to_income_ratio	1	661.1	204928	219878
respondent_id	50	8449.9	204878	211428

To check the significance of the variables in our model, we get the P-values and the result tells us that all the variables in the model are significant at alpha = 95%. The table containing the P-values has been attached in the appendix for reference.

Model Performance

The model performance is evaluated using the AUC curve and confusion matrix. The plots for the same can be seen below. The results used for the following plots use a threshold of 0.6





The interpretation of the results for Pennsylvania 2014 can be provided in terms of the odds ratio shown below.

Interpreting Results: PA 2014

- **Odds Ratio**

In logistic regression the odds ratio represents the constant effect of a predictor X, on the likelihood that one outcome will occur.

Predictor: Race and Ethnicity	**Coefficient**	**Odds**
White	0.7282939	107.15%
Asian	0.4509005	56.97 %
Hispanic or Latino	0.1865617	20.51 %
African American	0.1727634	18.85 %
Native Hawaiian	0.2847779	32.94 %
No Information Provided	0.3617244	43.58 %

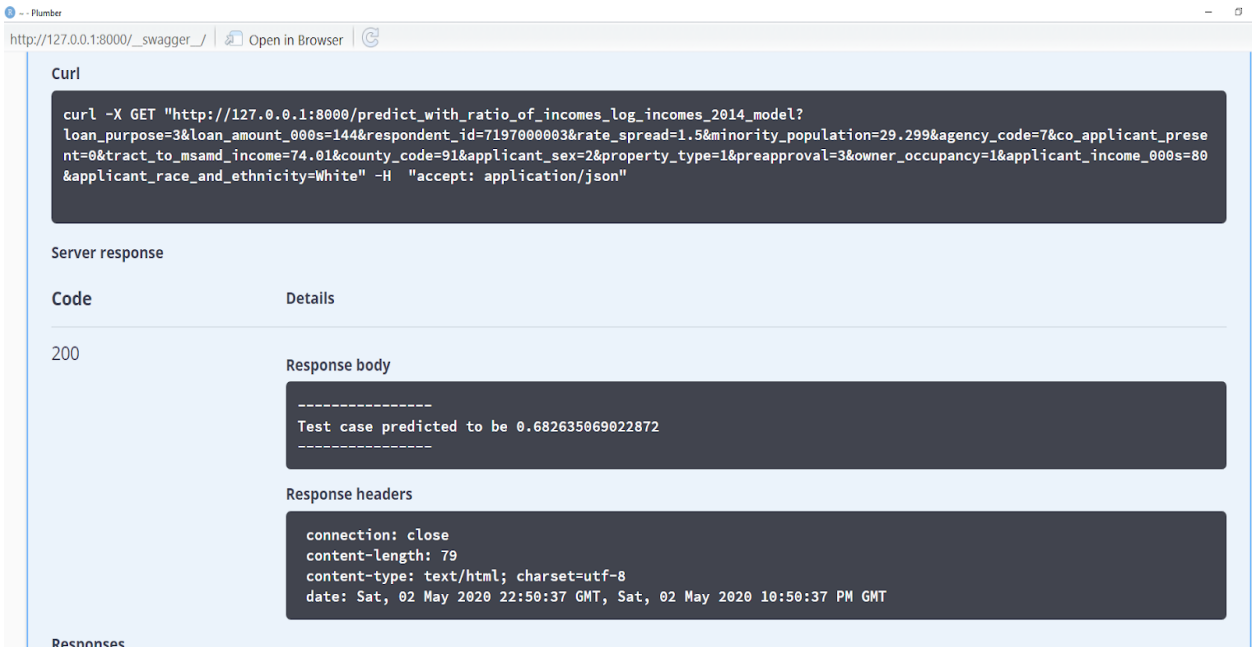
Looking at the table above, we can see that if you are African American applicant or from another minority, your odds for getting the loan are significantly low as compared to any White applicant. These results show that there could be a bias induced implicitly in terms of the race of applicant.

For details on other models please refer to the Appendix section.

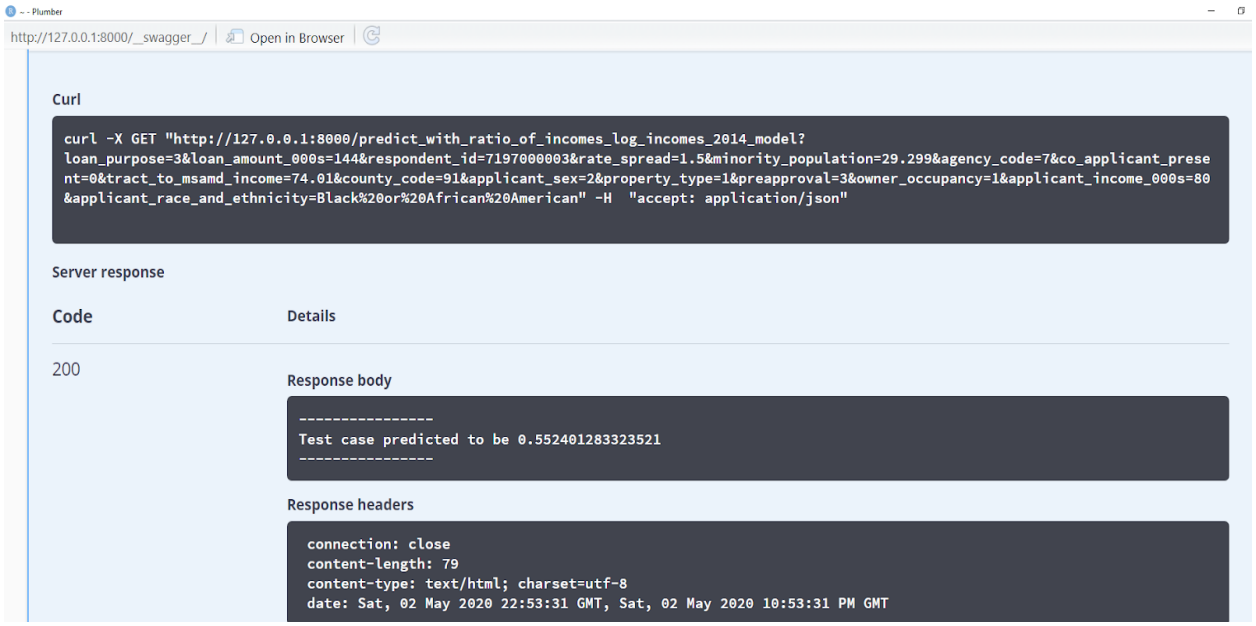
MODEL DEPLOYMENT

We used the R plumber library to deploy our models. To ensure that our models could be deployed, we first saved the models locally to disk. These models were then loaded in a R script. We then created R functions for running predictions using these models. These functions needed to be annotated for parameter details, etc as per the plumber library requirements. Once that was done, the plumber library does the rest. It creates REST APIs around the prediction functions and runs a local web server which listens on a specific HTTP port. Parameters and results are exchanged using JSON.

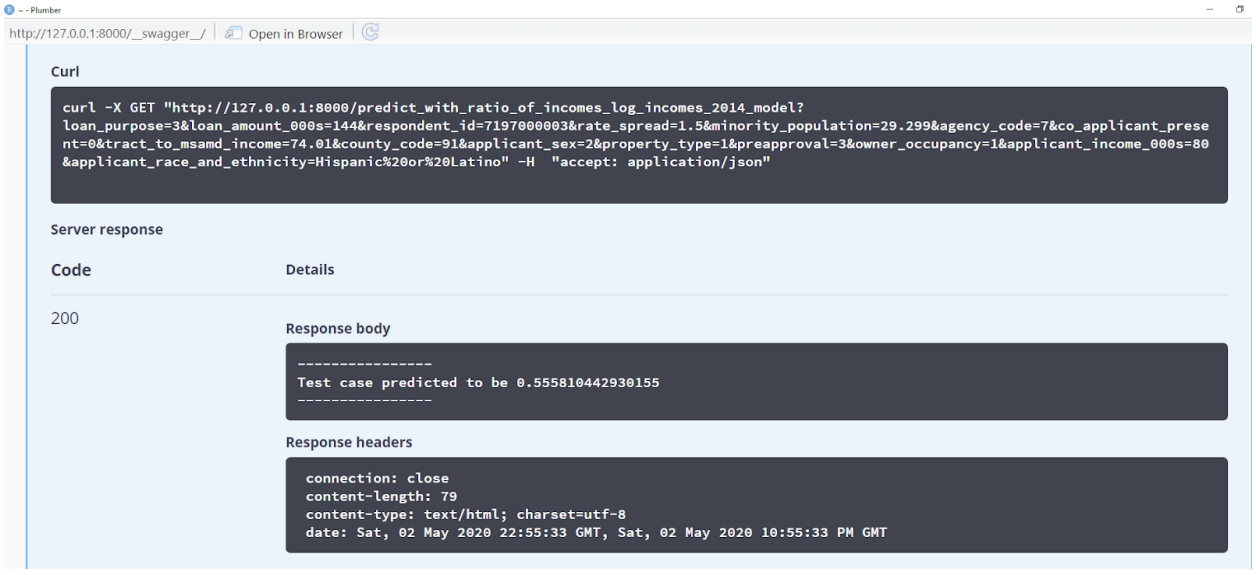
Prediction for a White applicant in Pennsylvania in the Montgomery County (county code 91)



Prediction for a Black or African American applicant in Pennsylvania in the Montgomery County (county code 91)



Prediction for a Hispanic applicant in Pennsylvania in the Montgomery County (county code 91)



The model shows that if we change the Race of an applicant from White to African American or Hispanic, that is, minority group, keeping all other values constant, we see a significant drop in the probability of getting the loan granted

CONCLUSION AND TAKEAWAYS

- **The Equal Credit Opportunity Act (ECOA 1974) prohibits race based discrimination.**

The Federal Trade Commission (FTC), the nation's consumer protection agency, enforces the Equal Credit Opportunity Act (ECOA), which prohibits creditdiscrimination on the basis of race, color, religion, national origin, sex, marital status, age, or because you get public assistance.

- **Data is telling a different story. Factors like debt to income ratio, credit scores, education level, bank balance, loan to value ratio, etc are missing in HMDA data**

Our models show some evidence of bias against minorities in Pennsylvania and Illinois. On a cautionary note, please note that banks use additional signals like debt to income ratio, credit scores, bank balance details, loan to value ratio, etc in their loan decision process. The HMDA dataset is missing this information. Please note that banks do provide this information in their filings with investing companies like corelogic, etc. The HMDA dataset, while extremely useful, is missing these key indicators.

- **Models can be improved with the above data**

Adding predictors like debt to income ratio, loan to value ratio, credit scores, etc to our model will improve its prediction power and along with it other factors like precision, recall, AUC curve, etc will improve.

- **Objective functions for models which have such a wide ranging influence on society cannot be binary. (Profit and loss), (Winners and losers), etc.**

Cathy O Neil in her book Weapons of Math Destruction talks at length about value based models for loan decisions, school teacher performance, resume scanning, etc. These models have the following things in common. Opacity, scale, feedback loop. Opacity is secrecy. People don't really understand math that well. There is an inherent assumption of an underlying order in math. Scale is the influence the decisions taken by these models have on society. Feedback loop. This is the most devastating of all. For the most part these models work on the underlying assumption that past performance is the sole predictor of future success. While that works well for things like industrial automation, self driving cars, etc, it does not translate well to overall society. People don't expect to be classified into buckets, etc. The long term effects of a loan not being granted, or not being able to get an interview due to race and gender are immense.

We need a societal debate about the utility of value based models like these. The Freedom of Information Act needs to apply to some of these models to ensure that their decisions can be audited and improved.

SOURCE CODE

The source code is available at this bitbucket repository

https://bitbucket.org/agiyengar/csp-571-02-final-project/src/master/

BIBLIOGRAPHY

[1] **HMDA API Documentation** - https://cfpb.github.io/api/hmda/fields.html
[2] **How to make your machine learning model available as an API with the plumber package**
https://www.shirin-glander.de/2018/01/plumber/
[3] **Community Reinvestment Act** - https://www.ffiec.gov/cra/
[4] **Equal Credit Opportunity Act**- https://www.justice.gov/crt/equal-credit-opportunity-act-3
[6] **Loan to Value Ratio** - https://en.wikipedia.org/wiki/Loan-to-value_ratio
[7] **Mortgage Lending Discrimination. Review of Existing Evidence** -
https://www.urban.org/sites/default/files/publication/66151/309090-Mortgage-Lending-Discrimination.PDF
[8] **Consumer-Lending Discrimination in the FinTech Era**-
https://faculty.haas.berkeley.edu/morse/research/papers/discrim.pdf
[9] **Big Data and the emerging ethical challenges** -
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5654190/
[10] **Census**
https://www2.census.gov/programs-surveys/popest/datasets/2010-2018/counties/asrh/
[11] **Visualizing Missing Values**
https://cran.r-project.org/web/packages/naniar/vignettes/naniar-visualisation.html
http://naniar.njtierney.com/articles/naniar-visualisation.html
https://stackoverflow.com/questions/17964513/subset-variables-in-data-frame-based-on-column-type
[12] **Extracting Specific Columns from Dataframe**-
https://stackoverflow.com/questions/10085806/extracting-specific-columns-from-a-data-frame
[13] **Stratified Sampling**-
https://stackoverflow.com/questions/20776887/stratified-splitting-the-data
[14] **McFadden's pseudo-R squared value**.
https://thestatsgeek.com/2014/02/08/r-squared-in-logistic-regression/
[15] **ROC Curve**-
https://hopstat.wordpress.com/2014/12/19/a-small-introduction-to-the-rocr-package/
[16] **Drawing Confusion Matrix**-
https://stackoverflow.com/questions/23891140/r-how-to-visualize-confusion-matrix-using-the-caret-package
[17] **Facet Wrap Function** - https://plot.ly/ggplot2/facet_wrap/

[18] **Significance Level of Factors-** https://stats.stackexchange.com/questions/100453/should-the-final-r-glm-include-only-significant-levels-of-factors
[19] **Variable Importances** - https://dataaspirant.com/2018/01/15/feature-selection-techniques-r
[20] **Correlation Matrix** https://towardsdatascience.com/how-to-create-a-correlation-matrix-with-too-many-variables-309cc0c0a57

APPENDIX

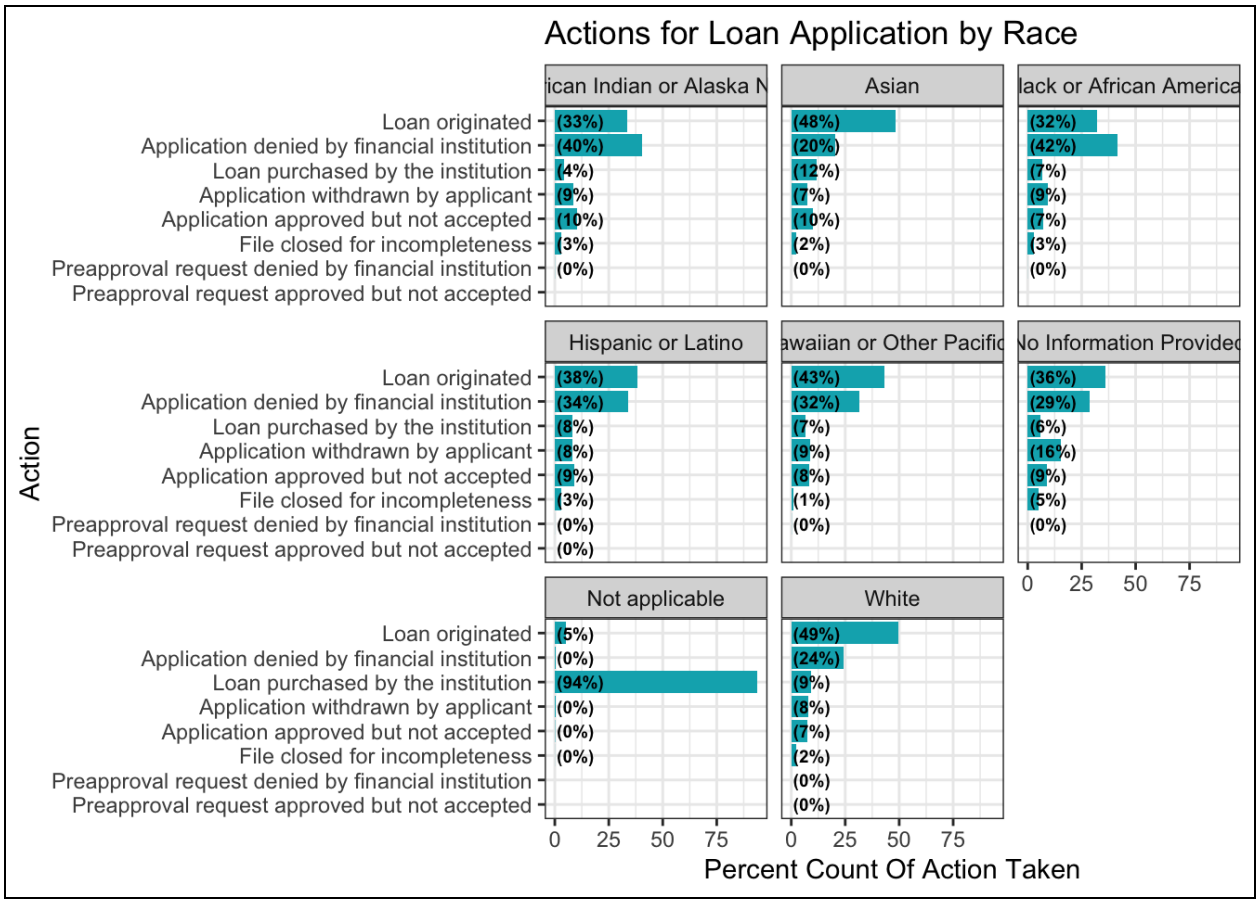
- DATA SOURCES

Descriptions about the attributes from HMDA data can be found here.

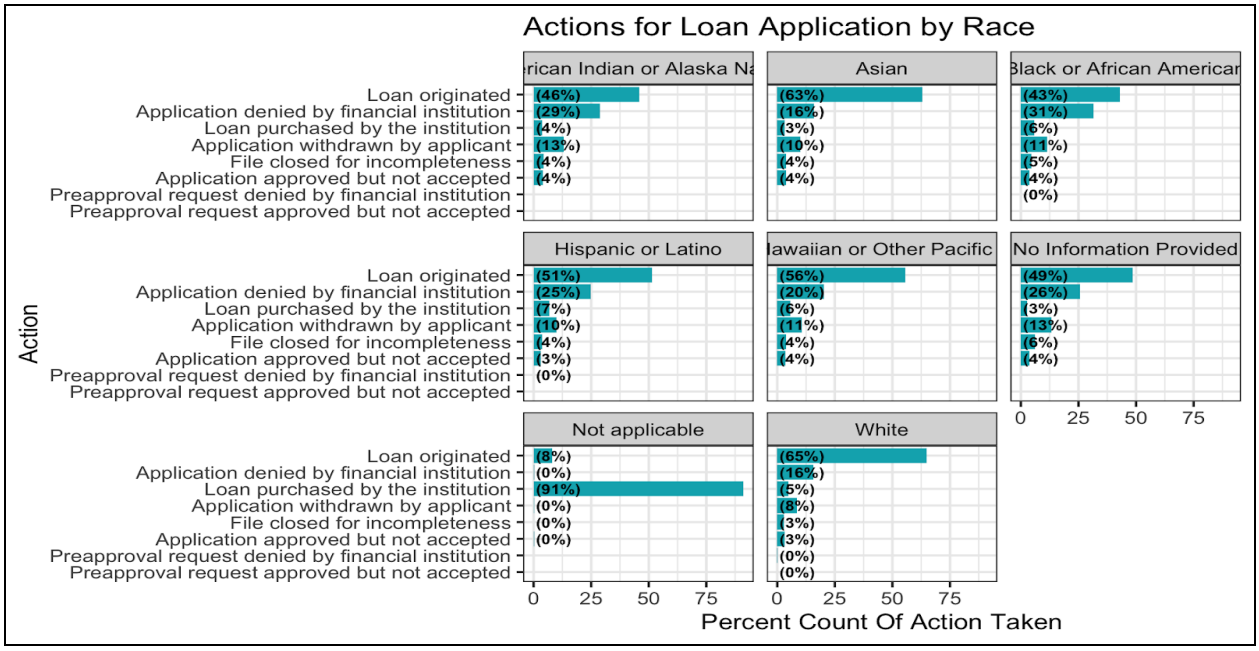
- EXPLORATORY DATA ANALYSIS

Follow are more plots showing action_taken distribution for PA and IL in different years

Action type according to race: PA 2007

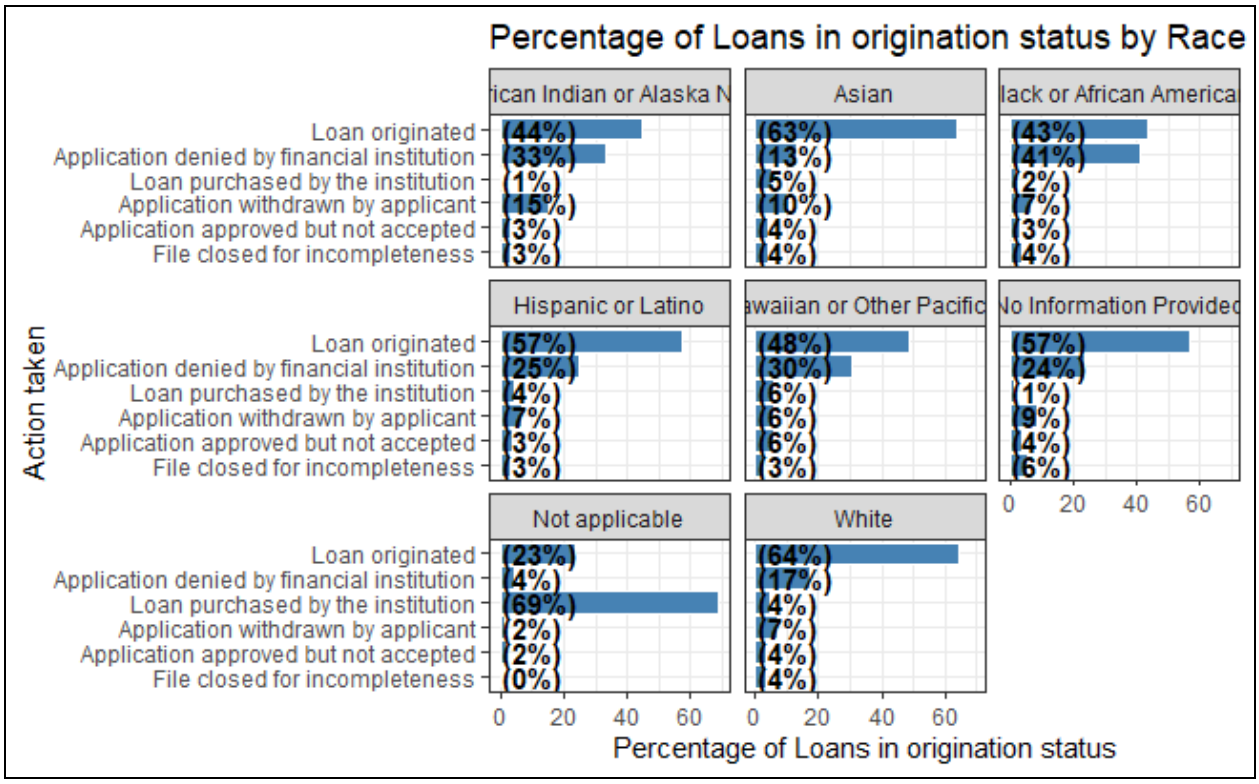


IL 2017

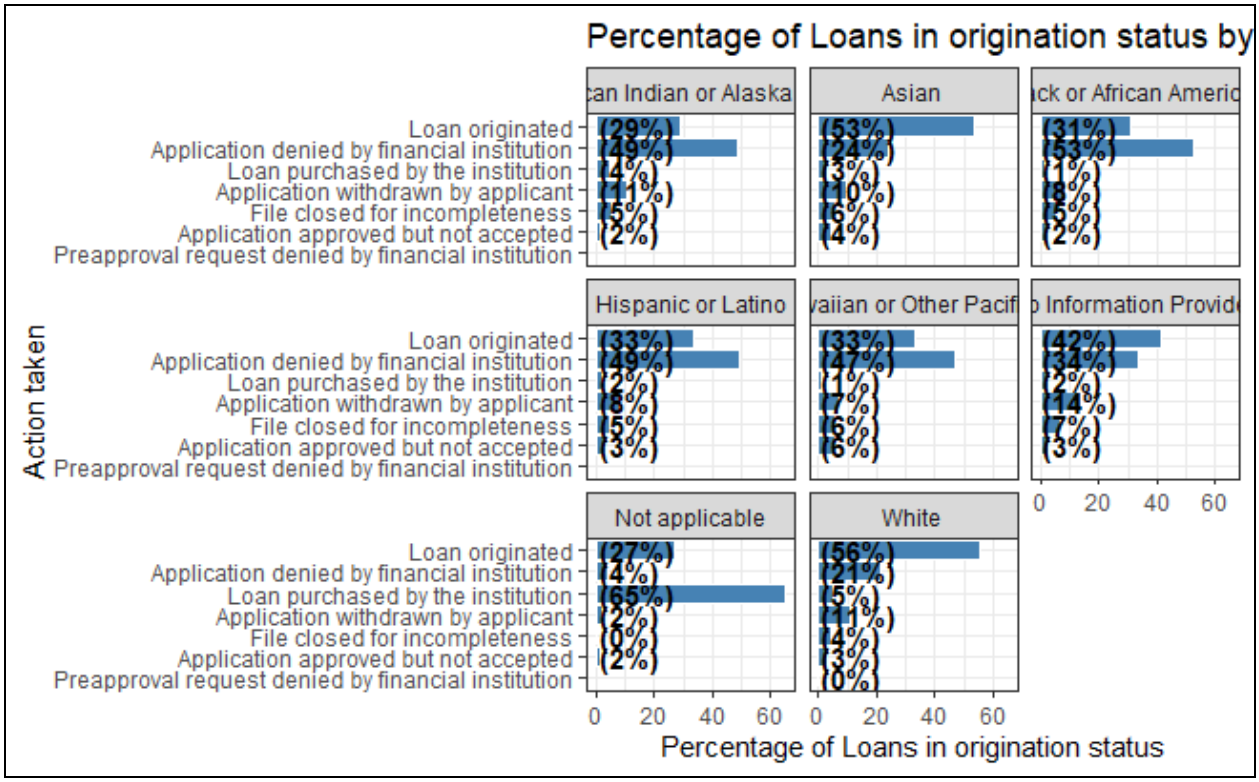


Same Operation performed of different counties in PA to see how the trend is per county.

Loan origination status distribution by race in Philadelphia county

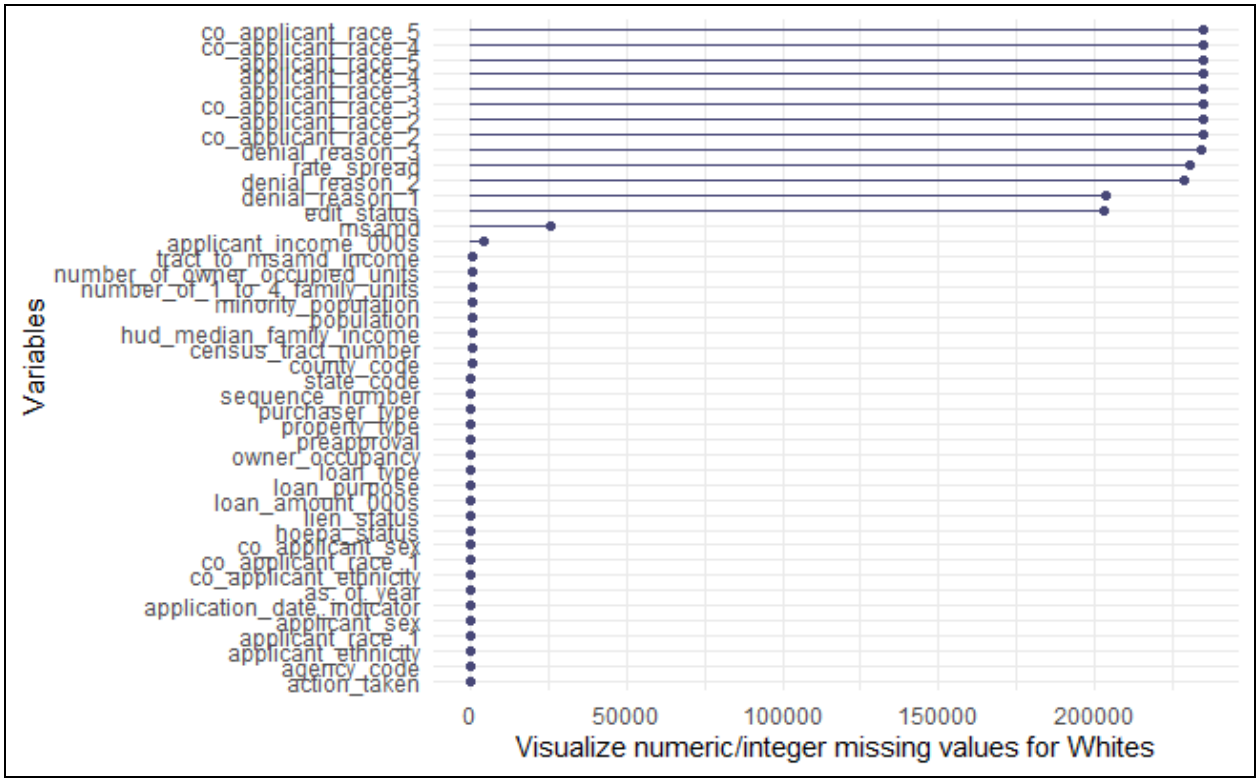


Loan origination status distribution by race in Allegheny county



- DATA CLEANING

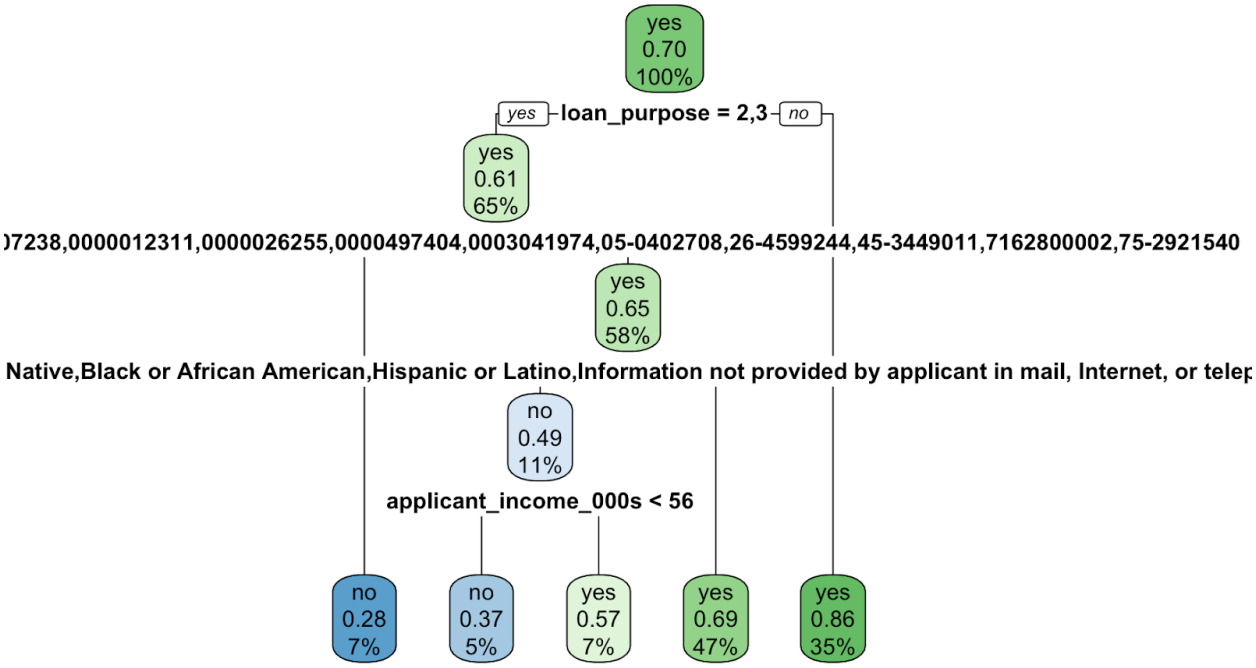
Missing Values Count Visualization



- MODELING

- **Decision Tree**

The following diagram shows the decision tree plot for 2014 PA. Here we can see that the model uses Applicant Race on level 2 to make decisions. This shows that applicant race is taken into consideration before income. This can be used as an evidence that the model could be biased according to applicant race.



- **Logistic Regression**

Additional Results for Logistic Regression models

- **P- Values for each predictor in the Logistic Regression Model**

Predictor	**P-Values**
applicant_race_and_ethnicity	0.00000000000000022 ***
owner_occupancy	0.00000002049 ***
preapproval	0.00000000000000022 ***
log_loan_amount_000s	0.00000000000000022 ***
applicant_sex	0.00000000000000022 ***
county_code	0.00000000000000022 ***
tract_to_msamd_income	0.00000000000000022 ***
co_applicant_present	0.00000000000000022 ***
agency_code	0.00000000000000022 ***
minority_population	0.00000000000000022 ***
loan_purpose	0.00000000000000022 ***
rate_spread	0.00000000000000022 ***

property_type	0.00000000000000022 ***
loan_to_income_ratio	0.00000000000000022 ***
respondent_id	0.00000000000000022 ***

Model details for PA 2008

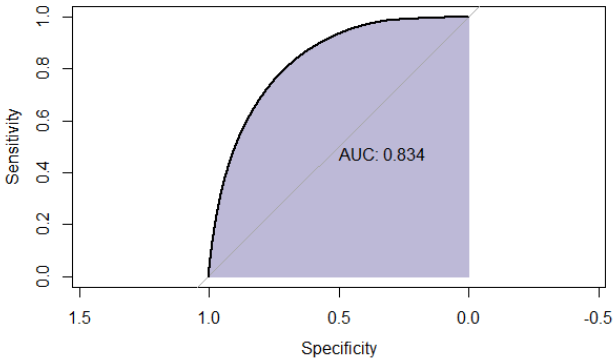
ANOVA

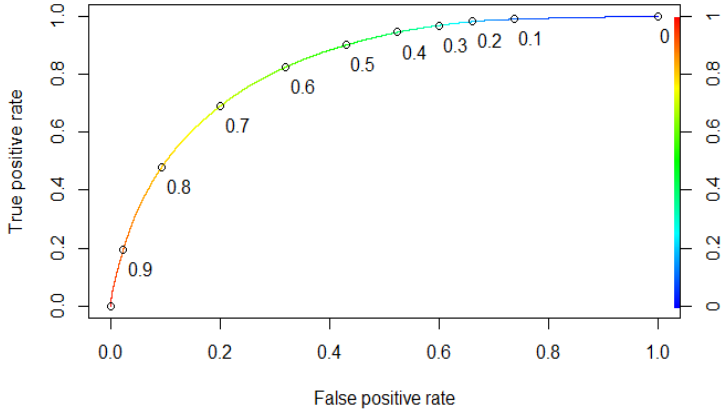
```
Anova for model
Analysis of Deviance Table

Model: binomial, link: logit

Response: loan_granted

Terms added sequentially (first to last)
```

	Df	Deviance	Resid. Df	Resid. Dev
NULL			416880	554488
applicant_race_and_ethnicity	7	11973	416873	542514
log_loan_amount_000s	1	1592	416872	540923
preapproval	2	18993	416870	521930
co_applicant_present	1	1938	416869	519992
county_code	25	3701	416844	516291
tract_to_msamd_income	1	4986	416843	511305
loan_purpose	2	13063	416841	498242
owner_occupancy	2	292	416839	497950
respondent_id	50	85828	416789	412121
minority_population	1	118	416788	412003
rate_spread	1	2724	416787	409280
agency_code	5	4347	416782	404933
applicant_sex	3	317	416779	404616
property_type	2	1792	416777	402823
loan_to_income_ratio	1	4137	416776	398686

```
NULL
applicant_race_and_ethnicity < 0.00000000000000022 ***
log_loan_amount_000s         < 0.00000000000000022 ***
preapproval                  < 0.00000000000000022 ***
co_applicant_present         < 0.00000000000000022 ***
county_code                  < 0.00000000000000022 ***
tract_to_msamd_income        < 0.00000000000000022 ***
loan_purpose                 < 0.00000000000000022 ***
owner_occupancy              < 0.00000000000000022 ***
respondent_id                < 0.00000000000000022 ***
minority_population          < 0.00000000000000022 ***
rate_spread                  < 0.00000000000000022 ***
agency_code                  < 0.00000000000000022 ***
applicant_sex                < 0.00000000000000022 ***
property_type                < 0.00000000000000022 ***
loan_to_income_ratio         < 0.00000000000000022 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Model Performance





Model details for PA 2014

1. **Logs of Applicant income and loan amount.**

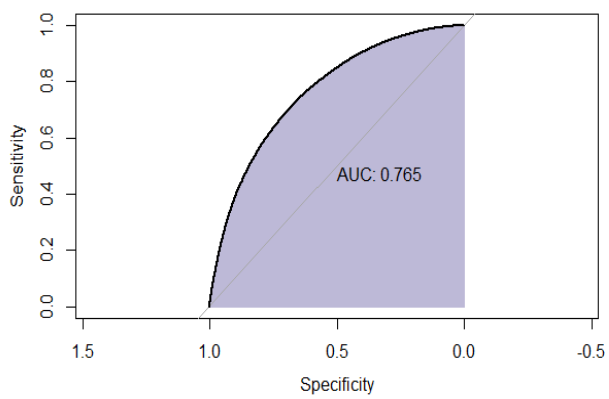
ANOVA

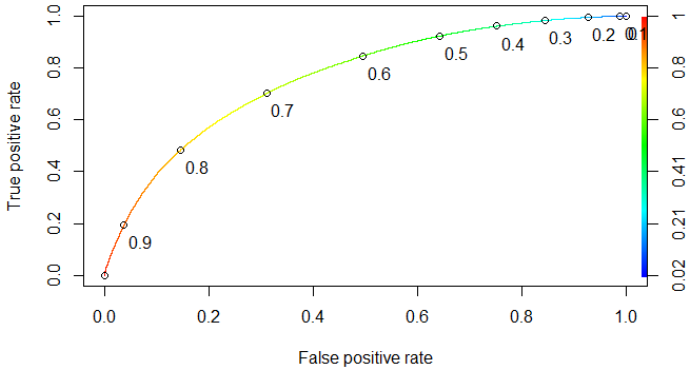
```
Number of Fisher Scoring iterations: 6

Anova for model
Analysis of Deviance Table

Model: binomial, link: logit

Response: loan_granted

Terms added sequentially (first to last)

                            Df Deviance Resid. Df Resid. Dev
NULL                                        204982     250576
applicant_race_and_ethnicity  7   7939.9    204975     242636
log_applicant_income_000s     1   6484.5    204974     236152
log_loan_amount_000s          1   1453.8    204973     234698
preapproval                   2   3002.6    204971     231695
co_applicant_present          1    338.1    204970     231357
county_code                  25    957.4    204945     230400
tract_to_msamd_income         1    279.0    204944     230121
loan_purpose                  2   7006.1    204942     223115
owner_occupancy               2    133.2    204940     222981
respondent_id                50  10568.3    204890     212413
minority_population           1    117.6    204889     212295
rate_spread                   1    336.2    204888     211959
agency_code                   5   1547.1    204883     210412
applicant_sex                 3     18.9    204880     210393
property_type                 2    372.5    204878     210021
```

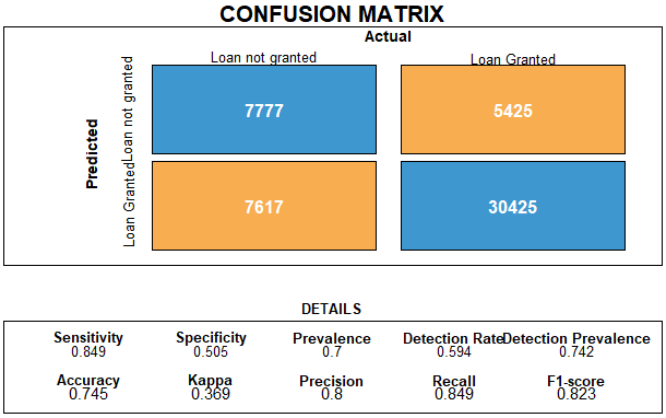
```
NULL
applicant_race_and_ethnicity < 0.00000000000000022 ***
log_applicant_income_000s    < 0.00000000000000022 ***
log_loan_amount_000s         < 0.00000000000000022 ***
preapproval                  < 0.00000000000000022 ***
co_applicant_present         < 0.00000000000000022 ***
county_code                  < 0.00000000000000022 ***
tract_to_msamd_income        < 0.00000000000000022 ***
loan_purpose                 < 0.00000000000000022 ***
owner_occupancy              < 0.00000000000000022 ***
respondent_id                < 0.00000000000000022 ***
minority_population          < 0.00000000000000022 ***
rate_spread                  < 0.00000000000000022 ***
agency_code                  < 0.00000000000000022 ***
applicant_sex                              0.0002927 ***
property_type                < 0.00000000000000022 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Mcfadden R Squared value
'log Lik.' 0.161848 (df=105)

Model Performance



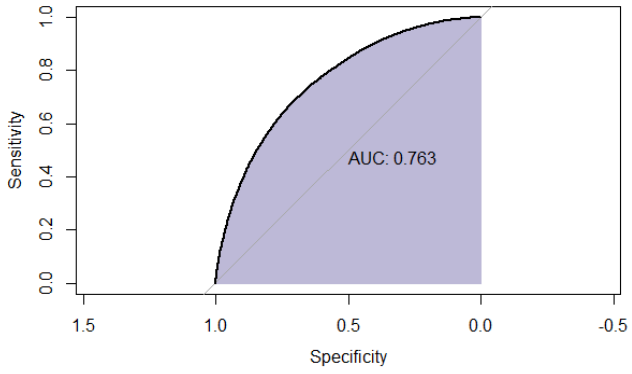


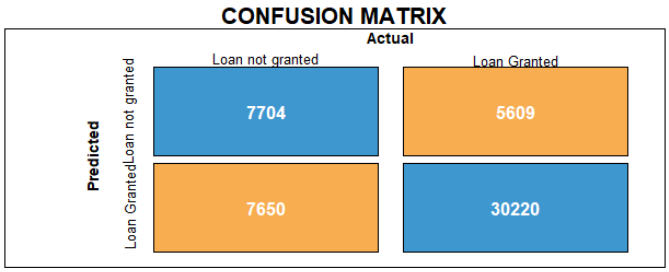


DETAILS

Sensitivity 0.849	Specificity 0.505	Prevalence 0.7	Detection Rate 0.594	Detection Prevalence 0.742
Accuracy 0.745	Kappa 0.369	Precision 0.8	Recall 0.849	F1-score 0.823

2. Including Census race based data.





DETAILS

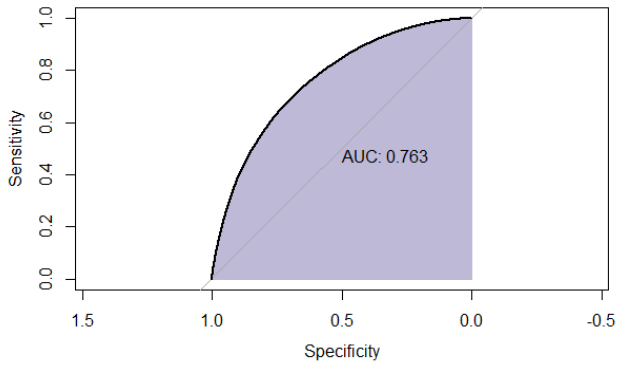
Sensitivity 0.843	Specificity 0.502	Prevalence 0.7	Detection Rate 0.59	Detection Prevalence 0.74
Accuracy 0.741	Kappa 0.359	Precision 0.798	Recall 0.843	F1-score 0.82

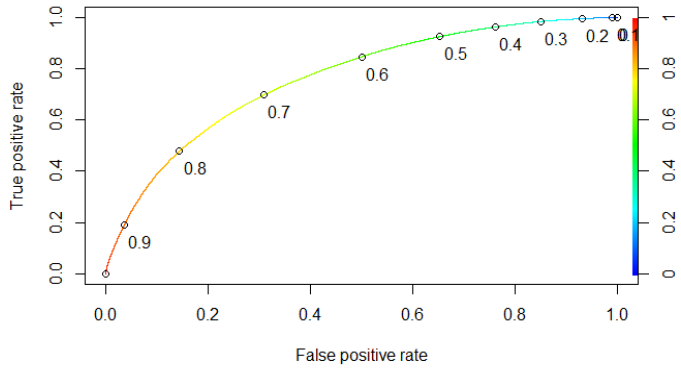
3. Including loan to home value ratio from Zillow median home values

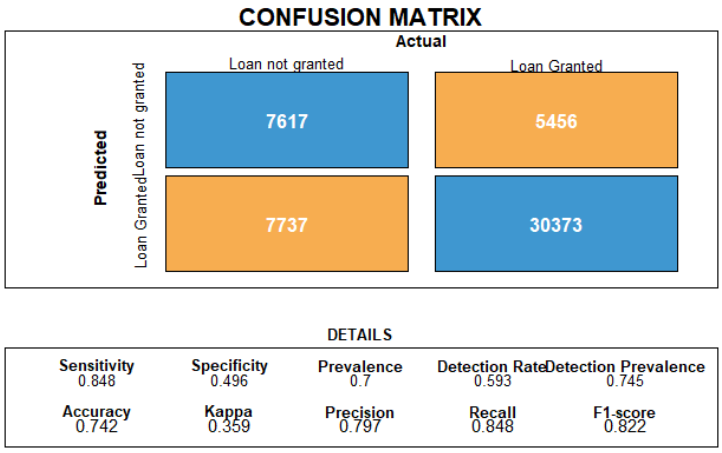
ANOVA

```
Analysis of Deviance Table

Model: binomial, link: logit

Response: loan_granted

Terms added sequentially (first to last)

                              Df Deviance Resid. Df Resid. Dev
NULL                                         204734     250125
applicant_race_and_ethnicity  7   7896.0     204727     242229
owner_occupancy               2     33.0     204725     242196
preapproval                   2   4139.7     204723     238056
log_loan_amount_000s          1   3571.6     204722     234484
applicant_sex                 3    210.0     204719     234274
county_code                  25    970.3     204694     233304
tract_to_msamd_income         1    751.4     204693     232553
co_applicant_present          1    909.5     204692     231643
agency_code                   5   3730.3     204687     227913
minority_population           1    118.3     204686     227795
loan_to_income_ratio          1   1073.7     204685     226721
loan_purpose                  2   6082.7     204683     220638
rate_spread                   1    466.7     204682     220171
property_type                 2   1091.4     204680     219080
loan_to_value_ratio           1     82.9     204679     218997
respondent_id                50   8478.8     204629     210518

NULL
applicant_race_and_ethnicity < 0.00000000000000022 ***
owner_occupancy                0.00000006886 ***
preapproval                  < 0.00000000000000022 ***
log_loan_amount_000s         < 0.00000000000000022 ***
applicant_sex                < 0.00000000000000022 ***
county_code                  < 0.00000000000000022 ***
tract_to_msamd_income        < 0.00000000000000022 ***
co_applicant_present         < 0.00000000000000022 ***
agency_code                  < 0.00000000000000022 ***
minority_population          < 0.00000000000000022 ***
loan_to_income_ratio         < 0.00000000000000022 ***
loan_purpose                 < 0.00000000000000022 ***
rate_spread                  < 0.00000000000000022 ***
property_type                < 0.00000000000000022 ***
loan_to_value_ratio          < 0.00000000000000022 ***
respondent_id                < 0.00000000000000022 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Model Performance







DETAILS

Sensitivity	Specificity	Prevalence	Detection Rate	Detection Prevalence
0.848	0.496	0.7	0.593	0.745

Accuracy	Kappa	Precision	Recall	F1-score
0.742	0.359	0.797	0.848	0.822

Model details for PA 2007

ANOVA

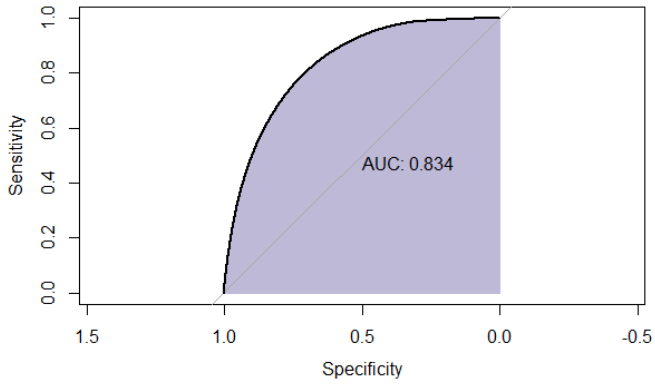
```
Anova for model
Analysis of Deviance Table

Model: binomial, link: logit

Response: loan_granted

Terms added sequentially (first to last)
```

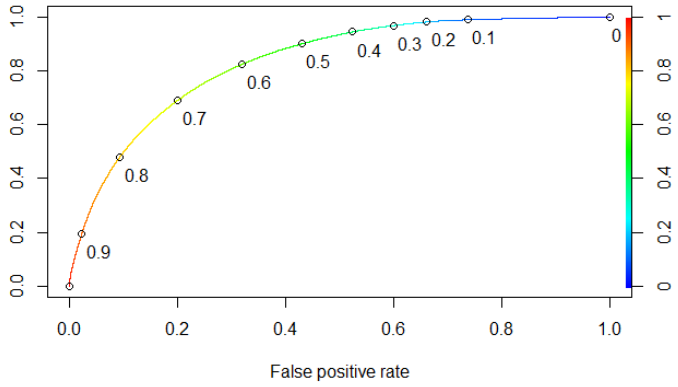
	Df	Deviance	Resid. Df	Resid. Dev
NULL			562189	752436
applicant_race_and_ethnicity	7	15314	562182	737122
log_loan_amount_000s	1	901	562181	736221
preapproval	2	29143	562179	707078
co_applicant_present	1	2815	562178	704263
county_code	25	7790	562153	696474
tract_to_msamd_income	1	6027	562152	690446
loan_purpose	2	17863	562150	672583
owner_occupancy	2	539	562148	672043
respondent_id	50	110261	562098	561783
minority_population	1	253	562097	561529
rate_spread	1	2175	562096	559354
agency_code	5	6963	562091	552391
applicant_sex	3	482	562088	551909
property_type	2	2886	562086	549024
loan_to_income_ratio	1	1949	562085	547075

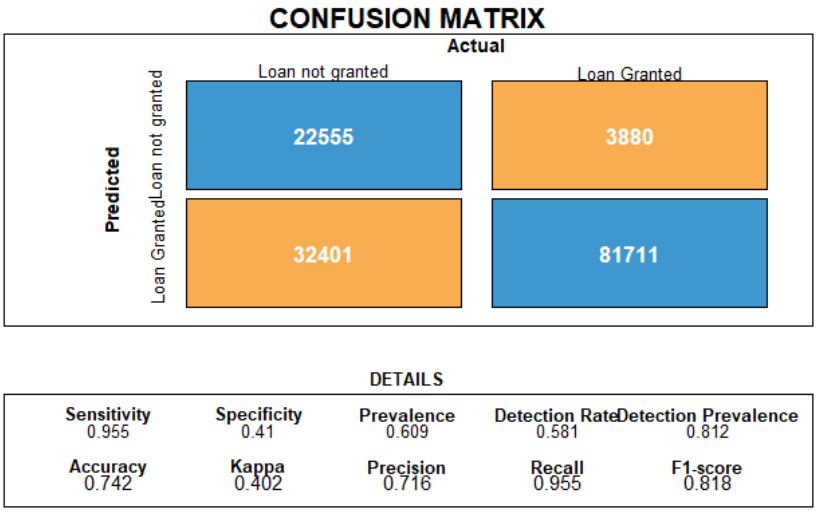
```
NULL
applicant_race_and_ethnicity < 0.00000000000000022 ***
log_loan_amount_000s         < 0.00000000000000022 ***
preapproval                  < 0.00000000000000022 ***
co_applicant_present         < 0.00000000000000022 ***
county_code                  < 0.00000000000000022 ***
tract_to_msamd_income        < 0.00000000000000022 ***
loan_purpose                 < 0.00000000000000022 ***
owner_occupancy              < 0.00000000000000022 ***
respondent_id                < 0.00000000000000022 ***
minority_population          < 0.00000000000000022 ***
rate_spread                  < 0.00000000000000022 ***
agency_code                  < 0.00000000000000022 ***
applicant_sex                < 0.00000000000000022 ***
property_type                < 0.00000000000000022 ***
loan_to_income_ratio         < 0.00000000000000022 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

We can see that the P values for the predictors chosen are all significant.

Model Performance







DETAILS

Sensitivity 0.955	Specificity 0.41	Prevalence 0.609	Detection Rate 0.581	Detection Prevalence 0.812
Accuracy 0.742	Kappa 0.402	Precision 0.716	Recall 0.955	F1-score 0.818

Please note that the above results are based on a threshold of 0.3 assuming that banks in 2007 were more willing to take risks and hence optimize for higher recall values.

Please look at the html outputs in the source code for details about other years.