Historical Surveys as Data

Project Background

The 1929 stock market crash devastated America's economy and triggered the beginning of a 10 year economic depression. During this time, American families were at risk of losing homes to foreclosure. To tackle the mortgage crisis, a federal loan program was created to refinance troubled residential homes and was titled the Home Owners' Loan Corporation (HOLC). HOLC created maps and area descriptions as part of a survey describing features of and threats to a particular area. Neighborhoods were graded based on the racial/ethnic presence, high and low-income families, and environmental problems. These surveys were used to determine whether families in these neighborhoods deserved loans or whether they did not; the decisions were largely based on race.

The dataset comes from crowdsourced transcription of digitized images from the National Archives and Records Administration. The NARA images cover over 100 American cities and neighborhoods. This project makes use of a subset of the entire database consisting of 172 observations and 39 variables. The University of Maryland Digital Curation Innovation Center (DCIC) is in the process of crowdsourcing the transcription of digitized survey images and the project focuses on cities for which corrected data is available.

Initial Thoughts

At project start, we developed several goals. The goals included:

- Identifying potential variables and stronger features;
- Normalizing and recoding selected variables as needed;
- Correlation testing of chosen variables for linear regression;
- Developing and fitting models (random forests, decision trees and support vector machines (SVM)) that accurately predicts the grading of a neighborhood based on economic, geographic and racial data; and
- Determining the priority columns for transcription correction for the project.

Process

This first step involved determining which columns would provide insight. The data was also split into an 80% training set and a 20% test set.

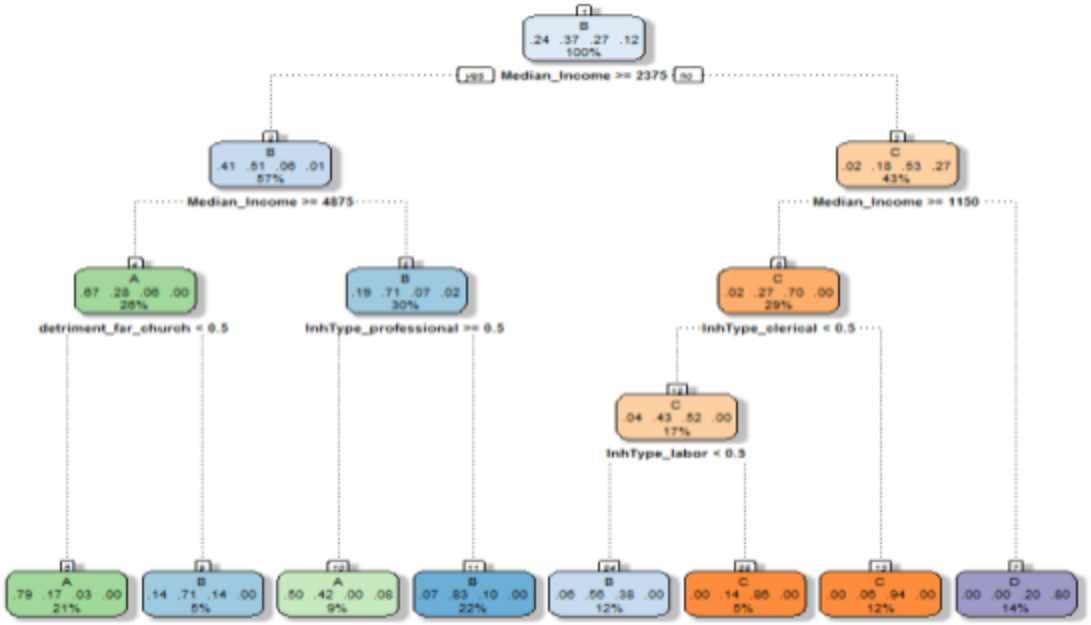


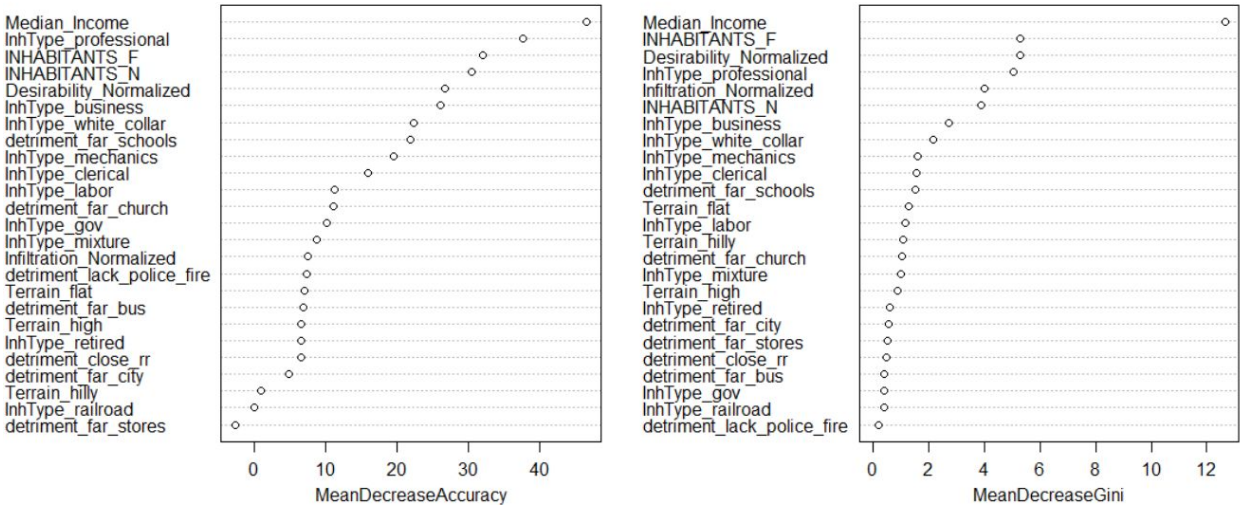
Fig. 1. Decision Tree for Mapping Inequality

The first model tested was the decision tree. In this model, median income was the most influential factor, with median income greater than $2,375 resulting in A/B grades. When state was included in the model, median income was the most influential feature followed by undesirable influences, specifically the distance from a school. In our new model, we removed state and the accuracy improved. With state, the accuracy was 67%. Without state, it was 70% and inhabitants type became more influential.

The second model tested was the support vector machine (SVM). Fitting was achieved by adding all variables and then removing sets of similar variables (such as inhabitant types and detrimental influences) and then individual variables and observing the effect on the prediction accuracy rate. The most influential variables proved to be:

- inhabitants who were laborers & wage workers;
- inhabitants who were professionals;
- median income;
- percentage of African Americans; and
- ethnic groups moving into area.

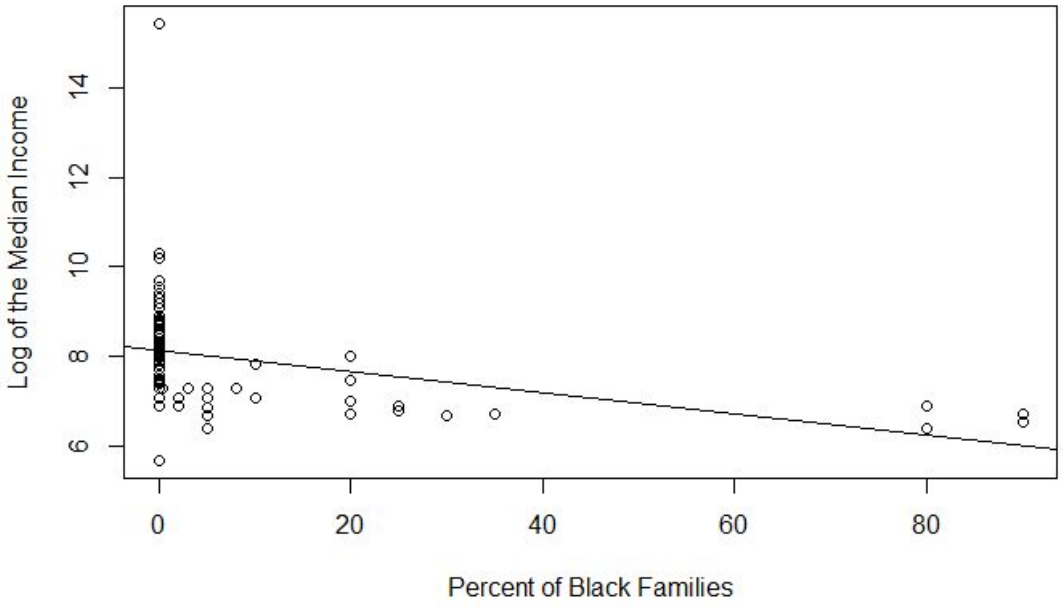
The SVM model was able to achieve an accuracy rate of 87.4%. However, accuracy varied drastically depending on the seed used, most likely due to the small number of observations in the dataset. In this model, the interplay of race, ethnicity, and class had more effect on prediction accuracy than the physical and social characteristics of the neighborhoods (terrain, detrimental influences).



The third model utilized was the random forest. This approach creates a large number of overfitted decision trees and then aggregates them by choosing variables by the highest accuracy and information gain. Our random forest model contained 2000 decision trees in total and yielded the highest accuracy of 96%. The model also arranged the explanatory variables in order of their importance.

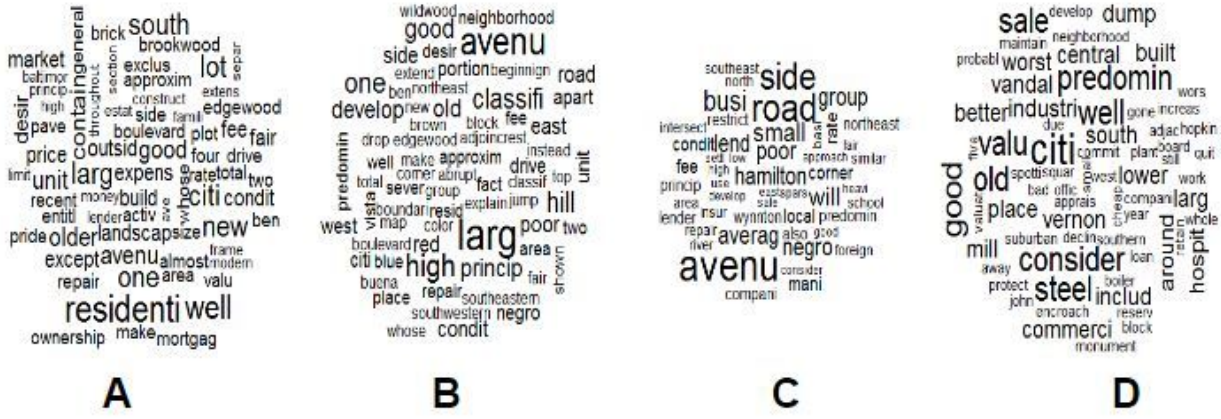
Utilizing all three models, it was clear that median income, the inhabitant's professional occupation, and their race greatly influenced the security grade. While the group hypothesized race would be the most influential feature, it was clear median income had more power.

Regardless, the group assumed that median income and race may be correlated and explored the relationship between race and median income within our data. We discovered a negative, linear relationship between the two features; the higher the percentage of black families, the lower the income in the neighborhood. In the end, the p-value was 2e-16. This result is akin to modern sociological and historical studies that show there is a pay gap between minorities and white professionals. Once more observations are available, the correlation should be retested for a more accurate linear relationship.



Information Gain

For the most part, the techniques were successful in normalizing the data and locating influential features. However, the group ran into difficulty with one particular column: remarks. The remarks column contained a large text block describing the physical location and the inhabitants that lived in the area. The group was unable to normalize this text block at it was too large. As a result, the group tested for inverse document frequency and regular DTM. Utilizing random forests, SVM, and decision tree, the highest accuracy was obtained using the random forest model with 96%. However, the models in general worked best when the remarks variable was not included. This was surprising as the group assumed the security grade could be influenced by the text block. To visualize the data from the remarks column, a word cloud was utilized.



The word cloud provides some insight into the Remark's contents. For instance, A & B ratings contain positive words such as desirable, well, and new. However, C & D ratings contain more negative words. There is a direct reference to African Americans in C. Other noticeable words include poor, vandal, dump, and small. While the remarks did not improve the model, the word cloud provided additional insight.

Conclusion

After testing several models, it was clear that the most influential features were

- median income;
- inhabitants type;
- percent of foreign born families living in the area; and
- percent of African American families living in the area.

Through testing, several lessons were learned. For one, the dataset must be larger to ensure consistency and accuracy. While high accuracy rates were generated with 172 observations, a more reliable model can be developed when all 6,000 observations are available for the analysis. Another lesson learned was that the model should start large and then be narrowed. While spliting the columns for normalization and descriptive statistics was helpful, testing small models were not as efficient as testing the whole model and removing insignificant variables.