STAT 495 Final Project

Tasheena Narraidoo, Jeff Lancaster, Luke Haggerty

DATA DESCRIPTION

The data we used was from the San Francisco Crime Classification Kaggle competition. There were 878,049 observations in the training set. The training set had 9 variables which broadly described the crimes, locations, and time occurrence. We manipulated the data to include separate variables for the hour of day the time took place as well as the month and year. We added an ID variable, which was helpful for our cross validation. Finally, for our ultimate model, we only used a sample of 200,000 observations (roughly a quarter of our train set), as using all 878,049 observations would have been too cumbersome.

EDA

After our data manipulation, we performed an extensive exploratory data analysis. We made use of the ggplot package in R to visualize the dataset in relation to the actual geography of San Francisco. We created several heat maps to isolate the the focal points of various types of crimes. We also developed a Shiny App that implemented a K Nearest Neighbors model to predict the locations of specific types of crimes based on the latitude and longitude of the crimes in the dataset. We created a bar graph to compare the overall frequencies of the different categories of crime. We also created several boxplots distributions to compare the frequencies of crimes based on time of day, day of week, month, and police department districts.

MODELS

We explored several options which included Random forest, knn and multinomial logistic regression for modeling our data. Our best two models were then a multinomial logistic regression model and a K Nearest Neighbors model. Based on the Kaggle scoring criteria, the multinomial model outperformed the K Nearest Neighbors model by far. Using a Multi-Class Log Loss scoring method, the multinomial model scored approximately 2.6, whereas the K Nearest Neighbors model scored approximately 26.7. To obtain our best multinomial model, we tried different multinomial models with different variable combinations, sample sizes, and iteration values (maxit) in order to get the best model with the best Kaggle values. We also ran Anova on our model to check the significance of our predictor variables. Our best predictors turned out to be Police District, Hour of the day, and weekday using a maxit of 200. After selecting the multinomial model as our final model, we wrote our own cross validation Multi-Class Log Loss to confirm the Kaggle score. Using a 5-fold cross validation, we arrived at a score of 2.8, which is pretty close to the 2.6 Kaggle score.

DIFFICULTIES

We experienced several difficulties over the course of this project. Firstly, the size of this dataset was difficult to deal with. With 878,049 observations, we were forced to take smaller samples of the data just to be able to run models without crashing R. Even with smaller samples, our code would often take a very long time to run. Furthermore, our options were limited by the fact that our response variable, Category, and most of our predictor variables were categorical variables. We had to find models that could accommodate categorical variables as both predictors and response variables.