**ISE 364 Final Project**

Technical Report

**Kevin Mittal**

12/12/19

**Introduction**

The purpose of this procedure was to create a classification model with one data set that could predict the value of a term on each tuple in a second data set. It was to take the input of an anonymized data set without column headers, explore and clean the data, and then to create an effective classifier that could be used on new data. The report describes the given set, the measures used to understand and organize it, the models created it from it, their evaluation, and the final results.

**The summary**

The aim of this project was pursued in initially cleaning up the data set by transforming categorical variables into numerical variables that could be understood by the models. Then, separating the data into a training and a test, various classification models including kNN, Decision Tree, and Random Forest, were trained and tested. They were assessed based on accuracy and precision. This experimentation revealed that the random forest classifier was the most effective model. This model was then used at the end to predict a new column of values on the second, futures.csv data set.

**Experimental Details**

The data.csv spreadsheet provided included 14 columns. Columns 0, 2, 4, 10, 11, and 12 all contained numerical values that needed no change for the classification models other than header names. Columns 1, 3, 5, 6, 7, 8, 9, and 13 on the other hand all contained categorical values that were anonymized by a character and a numerical value. There was also a final column, 14 that only included two values ‘SMALL’ and ‘LARGE’. There were also a number of tuples with the value ‘?’. This value represented a missing value and without other information about the relationship between various categorical options throughout each column, replacing the ‘?’ with a substitute numerical value such as 0 did not make logical sense.

**The Body**

Data Exploration and Cleaning

As expected, the most time consuming and difficult part of this experimental procedure was cleaning and organizing the provided data set. After first examining the types of values in each column, the final column 14 which consisted only of the categories ‘LARGE’ and ‘SMALL’ were easily mapped to values 1 and 0 respectively. Similarly, ‘G0’ and ‘G1’ were easily converted to binary 0, 1. However for the remaining categorical columns, it was observed that they all had the same format. Each tuple had a character, consistent throughout the entire column, such as A, B, or C followed a number. It then made logical sense to simply remove the character from each cell and turn the remaining number into an integer that could be read by the classification models.

Other strategies, such as creating a number of dummy variables for each of the unknown categorical columns were not pursued because of the added size they would have created in the data frame, and the resulting added computation time. The method chosen carried the risk of implying that A0 and A1 were closer than A0 and A5 thought that may have not been true.

The final stage of data cleaning involved accounting for the remaining empty or NULL cells, occupied by the ‘?’ character. In order to handle the ambiguity and potential error these values created, all rows containing the question mark were removed from the data set. Because these cells were missing a value, there was a risk in attempting to include the lacking rows, because the resulting model could be unfairly weighted. This process was done on both the provided data.csv and futures.csv for consistency. The final result of these efforts was a dataframe that could be used for classification models.

Model Evaluation

Multiple models were trained with the newly established and cleaned data set. Proceeding from basic analysis to more complex, kNN Decision Trees, and Random Forest models were all fit to the training set. After that a classification report was printed on the results of each model on the test set.

**Results and Discussions**

The kNN classifier resulted in a model that that was 72% accurate, which was as expected a base level of accuracy based on the relative complexity of the kNN algorithm. The Decision Tree Classifier then resulted in a 80% accurate model while the Random Forest was the most accurate with an accuracy score of 85%. Because of this the Random Forest model was used on the futures.csv data file to predict the column 14 values. This new data frame was exported into a csv file called final.csv.

**Conclusions**

Overall this experiment revealed that even when a large amount of data is anonymized, classification models can still create predictions that are fairly accurate with a little cleaning and preparation on input data set. Out of the 3 classifiers tested Random Forest was experimentally proven to be the most accurate, as expected from the most sophisticated of the classifiers chosen. The resulting model from data.csv had an 85% accuracy when tested, an accuracy that is reflected on the newly created final.csv file.

**Recommendations**

There are a couple sources of improvement in this experimental procedure. The two major weaknesses were the handling of the categorical variables and the handling of ‘?’ character in certain cells. In the case of categorical variables, it may have made more sense to use dummy variables instead for the categorical values. While this would take more time and add more complexity to analysis, it may have resulted in a more accurate model. Secondly replacement could have been attempted for missing cells, such as replacing each with a 0, and then results could have been examined to see the effect on accuracy and precision. This would have allowed more flexibility and adaptability on the futures.csv data set where some tuples had missing values.