DATA622: Homework 1 Essay

Eric Lehmphul

This homework required the selection of two datasets (one small dataset and one large dataset) and two to machine learning algorithms to predict a target variable from the given datasets. I elected to use the provided data source from https://excelbianalytics.com/wp/downloads-18-sample-csv-files-data-sets-for-testing-sales/ where I chose to use the 100 record dataset and the 50,000 record dataset.

From exploring the datasets, I noticed that the 100 record dataset contained categorical variables with very few records in each class. For example, Country had 76 unique values out of the 100 rows. Categorical variables that do not have enough instances in the data can be problematic for model building as they are only represented in either the training or testing dataset, not both. I ran into this issue when working with the small dataset. I overcame this issue by removing all variables that had severely underrepresented classes. The downside to doing this method is the loss of information gain from the other classes. The large dataset did not have the underrepresentation of classes. The bar plots that I generated showed the class distribution to be fairly uniform for most classes apart from Region.

The variable used from prediction in this assignment is the categorical variable, Order.Size. This variable was engineered using variables already present in the data. More specifically, Order.Size is the binned representation of the Units.Sold variable. I chose 5 distinct cutoff values to split the data: “0-1999”, “2000-3999”, “4000-5999”, “6000-7999”, “8000+”. This variable is ordered categorical variable with 5 levels, indicating that an Ordinal Logistic

Regression be used over a multinomial logistic regression. The second algorithm that used in the analysis was K Nearest Neighbors. The pros to using an ordinal logistic regression model is that it is much more interpretability compared to the KNN classifier, works on categorical data and is low computationally to produce. Some cons of ordinal logistic regression are that it makes assumptions about the underlying data, is prone to extreme outliers, and that it constructs a linear decision boundary. The pros to using KNN is that there is no training period, there are no assumptions made about the data, and is easy to implement. The cons of KNN are that it does not work great with small datasets, it can be computationally expensive, and is sensitive to irrelevant attributes.

The small dataset generated an ordinal logistic regression model with an overall accuracy of 0.25 and a KNN classifier with an overall accuracy of 0.6. The large dataset produced models with accuracies of 0.2023 for the ordinal logistic regression and 0.991 for the KNN classifier. I would suggest to business decision makers to use the KNN models due to the significantly higher accuracies if prediction is the only point of interest as the KNN model is considered a lazy learner.

The results echo the effect data size has on parametric and non-parametric algorithms. Parametric models are more constrained and do not need a large amount of data to be trained, though are usually outperformed by their non-parametric algorithms due to the inherent restrictiveness. Nonparametric data, on the other hand, is not restricted and can be very effective if there is an adequate amount of data to suffice the algorithm. KNN outperformed the Ordinal Logistic Regression models in both datasets. There is a clear improvement in the KNN algorithm when more data is present. This effect is likely due to the curse of dimensionality.