Dist. Machine Learning Report:

Durant Fullington and Preston Robertson

The project focused in on performing regression analysis on a dataset containing information related to used cars. The initial dataset contained 8,128 samples, and each sample contained 13 features, including attributes such as the name, selling prices, engine, fuel type, and transmission type. These different features all contained different scales and data types, where some represented integer or continuous values, and some are categorical and string variables. From these attributes, the project’s goal is to predict the price of a used car, based on the attributes of a car in the testing dataset.

1. Data Preprocessing and Feature Engineering

The First step to constructing our regression models was to perform data preprocessing, which included dataset cleaning and feature engineering. First, exploratory analysis was performed on the dataset, evaluating the type and importance of the attributes and clean-up the samples before using them to train and test our regressors. Next, the cardinality of the dataset was evaluated to find the unique values contained within each of the feature. The only feature removed was the ‘county’, which had 0 unique features. In addition, missing values contained in each feature were evaluated, and features which contained more than 10% of missing values were eliminated. These features included: ‘condition’, ’cylinders’, ’VIN’, ’Drive’, ’Size’, ’Type’, ’Paint’. This helps to clean up the dataset and minimize the amount of datapoints that have to be \_\_\_\_\_\_\_. The last step in the process for cleaning the data was to manually removes certain features that provided no direct benefit to the regression model. This included the Image, region, and website URLs, since all the data comes from one website, and the ID and craigslist post description. Ultimately, these values did not add-value to our models and could not be easily converted to numeric values to evaluate in our regression model. For Feature Engineering, the first step was to use a string indexer to convert the string attributes into numeric, categorical values. From here, Principal Component Analysis (PCA) was performed to reduce the number of features and resulted in having 3-Principal Components to represent 95% of the variance contained within the dataset. This dimensionality reduction helps the model to perform better and run faster than using a non-reduced dataset.

1. Regression Algorithm Implementations (Why we chose them)
   1. K-Nearest Regressor
   2. Simple Linear Regression
   3. Elastic Net Regression
   4. Random Forest Regression
   5. Support Vector Regression
2. Results and Discussion

The report will be broken into the following sections:

(1) Dataset Overview

(2) Preprocessing and Feature Engineering

(3) Regression Algorithms

(4) Results and Discussion

1. Dataset Overview:

The project focused in on performing regression analysis on a dataset containing information related to used cars. The initial dataset contained 8,128 samples, and each sample contained 13 features, including attributes such as the name, selling prices, engine, fuel type, and transmission type. These different features all contained different scales and data types, where some represented integer or continuous values, and some are categorical and string variables. From these attributes, the project’s goal is to predict the price of a used car, based on the attributes of a car in the testing dataset.

2. Preprocessing and Feature Engineering:

The First step to constructing our regression models was to perform data preprocessing, which included dataset cleaning and feature engineering. To start, an exploratory analysis was performed on the dataset to gather a better understanding of the characteristics of the dataset. This included evaluating the size of the dataset, different data types and what value the features provided. This helps to create a road map and general plan of attack on how to process the features and tailor the dataset for the regression models.

Next, the cardinality of the dataset was evaluated in order to find the unique values contained within each of the feature. Based on the dataset, the attributes all passed the cardinality check as they contained enough unique values to be either one-hot-encoded or left as they are to add value in the regression models. The Dataset was also checked for missing values and has a relatively low amount of missing data. No features were dropped based on the missing values threshold of 10%. Next, several of the categorical variables were one-hot-encoded for better use in the regression models. Finally, a string indexer was used to convert the string values into categorical variables for use in the regression models.

Two other important steps in the process for cleaning the data was to manually removes certain features that provided no direct benefit to the regression model, and to fill the missing values within the dataset. The variable we discarded of was the name of the car, as this variable is not very important for determine the price. For missing values, we used the pandas fillna method.

3. Regression:

(Discuss grid search)

2.1 K-Nearest Regressor

2.2 Simple Linear Regression

2.3 Elastic Net Regression

2.4 Random Forest Regression

2.5 Support Vector Regression

4. Results and Discussion:

The metrics used to evaluate performance was the R2 and Mean Absolute percentage error. These values allowed us to compare the performance of the different regressors with each other and showed that the KNN and Random Forest regressors performed the best, as both had the highest R2 Scores and Lowest MAPE scores. (See Bar plots)

The MAPE metric was chosen because it calculates the error of the model, no matter what scale the target variable is. This is since MAPE calculates relative percentage error with respect to actual output. This means that models with smaller values perform better, with less error between the predicted values and the true values of the used car's selling price.

The R2 Metric was used to evaluate the amount of variance that has been explained by the independent variables in the model. This means that models with higher values better fit the data, as they were able to capture more of the variance within the dataset.

In addition, the testing and training scores were used to evaluate how the models fit the data. None of the models seemed to overfit the data too badly, as the training and test scores were very close to each other, for every test.

Overall, using KNN and Random Forest would allow for decently accurate prediction of a car's price, based on the features contained within this dataset.