**Introduction: EDA and Prediction of Used Car Prices on Craigslist**

The used car market plays a vital role in the global economy, providing affordable transportation options to millions of consumers while also serving as a key economic indicator of the US economy. Unlike the new car market, which is heavily influenced by manufacturing trends and corporate strategies, the used car market reflects real-world consumer preferences, behaviors, and financial constraints. Additionally, used cars contribute to a greater sustainability of manufactured goods, reducing waste and prolonging the lifespan of vehicles. It enables individuals access to vehicles for often significantly cheaper prices than newly-created cars.

In the exchange of used vehicles from consumer to consumer, the market contains formal and informal sellers. For example, Carmax is one of the largest used car retailers in the United States that facilitates transactions between buyers and sellers. However, in this project, the pricing of informal sellers is of greater interest. On Craigslist, a classified advertisement site, used vehicles take up a significant portion of advertised products. The pricing of used cars on the site could be interesting to analyze, as it would provide insight into the features that users and buyers value the most in a car. Understanding consumer preferences is important to tracking price changes in the more volatile used car market, as well as for sellers to gauge the true value of their vehicle.

Following the goal of predicting used car prices, I sought to find a dataset of used cars, including the price that the car demanded, as well important features such as odometer value, age of the vehicle, color, model, etc. Using such a dataset, I could then use a predictive regression model to predict used car price based on the description of a car.

**Data Description**

I chose a dataset from Kaggle that was scraped off of the Craigslist website listings. The data reflects used car listings from 1999 to early 2023. The columns were created based on the information reflected in the listings, namely information about the year, odometer, region, price, manufacturer, model, condition, size, type, paint-color, cylinders, etc. Based on these columns, I chose 12, including 1 dependent variable of price, that best captured the potential independent factors that influenced the price. I removed columns such as region and state because it was irrelevant to the consumer preference prediction I was aiming to conduct. Additionally, some features contained many null/missing values that made it difficult to include in the cleaned dataset.

After selecting the appropriate columns, I removed outlying values in the odometer, year, and price columns. Additionally, model and manufacturer had many distinct values that would make it difficult when encoding the categorical variables. Therefore, I instead chose the first 50 and 20 most often occurrences of model and manufacturer, respectively, and grouped the remaining into an “other” category. At this point, the dataset is cleaned of null values and processed to optimize for regression models.

**Exploratory Data Analysis**:

Before diving into the predictive models, it is important to overview potential trends and correlations in the data. In the first histogram plot of price counts, it is notable that the curve peaks within the $5,000 to $10,000 range, meaning most used cars fall within this price range. Since the graph is right skewed, there are much fewer cars with prices exceeding $20,000. Cars priced under $5,000 are also

frequent, suggesting a significant number of budget-friendly options or older vehicles, rather than expensive or luxury cars. This may relate to Craigslist's reputation for unreliability and susceptibility to fraud, which makes listing expensive cars less incentivized.

The second graph is a scatter plot, depicting the correlation between odometer values and price. Since the odometer measures the distance traveled by the car, in miles in this case, we expect that the higher the odometer value, the less likely it is for the car to be priced high. This is because a car with many miles traveled is expected to last a shorter time than a car with fewer miles traveled. The scatter plot reaffirms this logical trend, as a few points with high odometer readings are also listed for a high price. Additionally, however, we do not see this trend occurring the other way around. For cars with low odometer readings, the price listing can still be high or low, indicating that there are other factors that significantly correlate with how used cars are priced.

Another potential factor that plays into pricing is the overall condition of the car. Craigslist has 6 main categories for car condition, from new to salvage. By examining the boxplot figure of each condition and the price, we see that cars in the "fair" and "salvage" conditions have a significantly lower median price at below $5000, while other conditions fall close to $10,000. The disparity between the different conditions signals that a worse used car condition can significantly impair the value consumers assign to a car.

There are additional factors that stem from the fact that the dataset came from Craigslist. In the histogram of yearly counts of listings, we see a steady increase from 2000-2013, then a steeper decline after 2013. These trends can partially be attributed to the loss in popularity of Craigslist as an advertising site. Instead, commercially used car dealers like CarMax and Carvana have gained in popularity, and also eBay has largely replaced Craigslist when it comes to selling goods online. Another note in the trend is that in 2009, shortly after the 2008 financial crisis, we saw a sharp dip in used car listings. This is an example of the used car market being an indicator of consumer behavior. During the recession, people generally consume fewer products, including the purchase of used cars.

Another question regards consumer preferences about the type of car. Based on the boxplot, we notice that certain types of cars have a higher price, such as pickup trucks and trucks in general. This may have to do with the fact that Craigslist has more male users than females, and truck purchasers are ⅔ males. Increasing demand may correlate with a higher price.

Most cars sold on Craigslist are gas-fueled. One reason is that used cars reflect past manufacturing decisions. Hybrid and electric vehicles historically were not purchased in significant quantities by the public, therefore mostly gas cars were manufactured, and now these gas cars take up a significant portion of the sales of used cars. The same applies to why automatic-style cars are prevalent in the used-car market.

When it comes to the paint color of a car, certain colors, like black, white, and unique colorings, were more popular amongst buyers and thus were most prevalent. Other colors seemed less appealing, such as purple and green. Another reason may be that there are not enough data points for cars of these colors to accurately represent the true median price.

Looking at numerical correlations through a heatmap, we notice that there seem to be significant correlations between the different variables, which poses a concern over multicollinearity.

**Preprocessing**

Given that there were many independent variables to gauge against the dependent variable of price, I wanted to test the accuracy of 4 different models: multiple linear regression, decision tree, random forest, and k-nearest neighbors. I established my baseline to be the mean price of the data points. This baseline would allow me to assess the improvements that the models are able to yield.

Before training and testing the model, I preprocessed the data, encoded the categorical variables, and scaled the numerical variables. Since the "condition" column has a clear ordered progression, I also chose ordinal encoding to translate the "condition" column into numerical data that the regression models could handle.

I split my data into an 80-20 train test set. Since my dataset contained 120 thousand rows, I chose to sample 30 thousand points instead of using the entire 120 thousand data points. This would allow it to still maintain accuracy, while also reducing the computation time for each model.

**Models and Interpretations: Understanding the Results and the Reasons**

My first model was on multiple linear regression. I obtained an r-squared value of 70%. This r-squared value signals a fairly strong correlation between the predicted test set and the actual prices of the test set. Compared to the baseline, this model was able to improve prediction with greater accuracy than using the mean price, with less deviation from the actual. To further tune the model, I added the polynomial features function, which generates new features that are a combination of the existing ones. This can help capture non-linearity in the data so that the model can approximate more complex, curved relationships. A drawback to increasing model complexity is the potential for overfitting. However, since I used a moderate-degree polynomial model of degree 2, it can capture the true trend without fitting noise. Using the polynomial features in the model, I got an r-squared value of 77.6%, which is an improvement from the linear regression model. Looking at a few features, that were prevalent in the polynomial features model, year and gas, year and clean condition, odometer and gas, etc, were important in the model. This means that interactions, when taken into consideration, yield a more predictive model.

The second model I tested was the K nearest neighbor (KNN) model. I used GridSearchCV during hyperparameter tuning to search through a predefined set of parameters for a model with the highest performance. The GridSearch yielded an n nearest neighbors of 15, where the 15 closest points are considered in the prediction model. This model performed better than the multiple linear regression model, with an r-squared of 75.7%. The reason for this is that the linear regression model likely exhibits underfitting, which the KNN model accounts for. However, it performed worse than the polynomial features model because the KNN model might not capture global trends as well. After all, it inherently looks only at local trends.

The third model I tested was a decision tree model, with an r-squared of 58.7%. Potential reason for why the r-squared value was so low was that a decision tree tends to learn the training data so well that it becomes prone to overfitting, and will not generalize well to unseen data in the test set. Moreover, while decision trees can capture nonlinear relationships, they do so in a piecewise fashion, which makes it more difficult for the model to perceive smooth trends in the data. A last potential reason is that decision trees work best on data that naturally divides into clear, well-defined partitions. If the data does not have obvious groupings, decision trees may struggle to produce accurate predictions.

The last model that I tested was a random forest model, which is an ensemble method of individual decision tree models. This model had an r-squared of 79.15%, the highest of the models. Comparative to decision trees, combining into an ensemble in the random forest model helps reduce

variance, leading to better generalization and improved performance on the testing data. Decision trees by themselves are prone to overfitting, but by averaging the predictions of multiple trees, random forest significantly reduces overfitting while still being flexible enough to capture complex patterns in the data. Moreover, unlike the linear regression model, the random forest has high flexibility because it does not make strong assumptions about the data.

Based on the final random forest model, I chose to train then test the model on the entire set of data, with the same 80-20 split. I wanted to understand how sampling affected the accuracy of my model. Ultimately, I got an r-squared value of 82.2%.

**Conclusions, Limitations, and Applications:**

In summary, the model that yielded the highest accuracy was the Random Forest model because it was able to avoid overfitting by taking into account multiple decision trees, and also because it can handle complex, non-linear correlations between the data. Based on this, I was able to conclude that the pricing of used cars on Craigslist has significant correlations with the features of the car. The most important features were the listed year and the odometer rating. Even in the polynomial features model, these two features consistently played a large role in the high accuracy of the model. This makes sense because consumers value the longevity of a used car. These two metrics are largely correlated with how long a used car is expected to last.

Potential limitations to the model include the following:

1. **High occurrence of missing values**: After cleaning and selecting columns and data based on completeness of data, there is a possibility that important trends are filtered out due to the selection process.
2. **Craigslist as a source of data**: Data taken from Craigslist may not be fully reflective of consumers based on the demographics of people that use the site. More specifically, the age range of users fall between mid-40s and late 50s, which is a result of the history connected with the site that generated continued loyalty from these groups. Therefore, the model reflects these specific consumer segments, rather than other demographic segments.
3. **Macro Trends**: The predictive does not account for the market changes the influence price. As seen in the EDA of prices over time, economic events such as inflation or recessions impact the demand, and thus the price, of used cars.
4. **Overfitting and Underfitting**: The Random Forest model is still susceptible with over and underfitting. If the model is too complex, it may memorize training data rather than generalizing to unseen data, leading to inaccurate predictions. Conversely, if the model is too simple, it may fail to capture important patterns in the data, resulting in poor performance.
5. **Correlation between independent variables**: There is a potential limitation that may arise of multicollinearity. When two independent variables exhibit a correlation, separate from the dependent variable, this correlation may skew the results of the predictive model, and not accurately portray feature importance. For example, if “model” and “fuel” are highly correlated, the weight of “fuel” may be underestimated because it could be partially captured in the weight of “model” in the final predictive model.
6. **Limited columns**: There is information that is not captured in the dataset that would be useful to consider. It would have been better to understand the initial price of the car, before it became a

used car. Similarly, understanding whether or not the car had actually been sold would also be insightful. These two features would have been a better gauge of consumer preferences

The information gained from this predictive model could be utilized in the following ways:

1. Car dealerships could use the model to determine what price consumers value certain cars based on their features. By predicting prices, dealerships can then make better purchasing decisions, ensuring that they buy vehicles at a price that allows for a reasonable profit margin.
2. For individual buyers and sellers, Individual car owners can use the model to estimate the selling price of their vehicle based on its current condition and features, ensuring they receive a fair price. On the other hand, buyers can evaluate whether the asking price for a car is reasonable and use the model's output as a negotiation tool.
3. Insurance companies can use the predicted car price to calculate accurate premiums. For example, higher-value cars might require higher premiums, while older or lower-value cars might qualify for lower rates.