Project Final Report

Pre-Owned Cars Price Valuation

Final Presentation Video

Introduction and Background

The used car market offers competitive prices based on factors like make/model, year, mileage, condition, and location. However, consumers often lack insight into whether a car is fairly priced or why it's priced a certain way.

While there has been significant research done on predicting car prices, the raw prices alone provide little value to consumers. The key questions are: Is this price fair? What factors influence it? Are these factors important to the buyer?

We aim to simplify the decision-making process for consumers by providing a fair-price estimate for a used car and a breakdown of key features influencing the price. By offering transparency into what contributes to the price, we empower users to select the car's features that best suit their needs and budget.

Dataset Description

The dataset is derived from https://www.cars.com/ containing 4,000+ data points of various used car listings. For each data point, there are nine distinct features: Brand & Model, Model Year, Mileage, Fuel Type, Engine Type, Transmission, Exterior and Interior Colors, Accident History, Clean Title, and Price.

Dataset Link

Used Car Dataset

Problem Definition

Given the various features of a car, such as brand, model, mileage, etc., our project aims to predict the fair price of the car. Additionally, we aim to classify the deal as a "bad," "fair," or "great" deal. We also seek to identify features having the largest impact on price which will provide more explainable results consumers can use to make informed decisions about whether the price reflects the value of the car based on features they care about.

Project Goals

The primary goal of the project is to build a model that predicts whether a used car is a good deal based on its attributes. We will ensure that resources on the clusters, including computing power and storage, are used sustainably and that data is sourced ethically (obtaining necessary permissions).

Methods

Pre-processing methods

We went through many iterations of data preprocessing to land at out finalized dataset. One of our main challenges was encoding categorical columns like brand, model, exterior color, and interior color. Intially sought to utilize one-hot encoding to break down these values. This proved to be a poor strategy, as some columns had very high dimensionality, thus, creating a lot of columns. To combat this, we tried consolidating categories into various buckets, then using one-hot encoding. This reduced dimensionality down to roughly 112 columns. However, when running our model with this dataset, our loss was extremly high. In our final iteration, we tried other encoding methods. For example we utilized fequency encoding for exterior and interior color. This allowed for unique colors to be represented as "rare" while more common colors (black, gray, sliver, etc) were weighted a "common." Addditionally, we used ordinal encoding for fuel types to capture the different types without increasing data dimensioality. Outside of these categorical encoding challenges, we did introduce new features. Engine was broken down into three columsn: cylinder count, displacment, and horsepower. This helped us break down the descriptions in the engine column into more quantitative values to help our model learn. Model year was converted to car age. This was necessary to calculate the mileage per year, giving us a representation of how much wear and tear was put onto a car in a year.

Machine Learning Models

1. Artificial Neural Networks (Supervised): Neural networks in deep learning-based approaches fit non-linear and complex correlations between the attributes. We attempt to capture these by building a Multi-Layer Perceptron classifier.

Scikit Link

2. K-Means clustering (Unsupervised): We will use the K-Means heuristic algorithm to cluster the data points into categories. This is effective since similar data points will be grouped in the same cluster. This pattern will help us identify similar used cars and bucket them into the same category.

Scikit Link

3. Random Forest (Supervised): To prevent overfitting of the model due to a fixed number of points, we will implement the random forest model to classify the deals.

Scikit Link

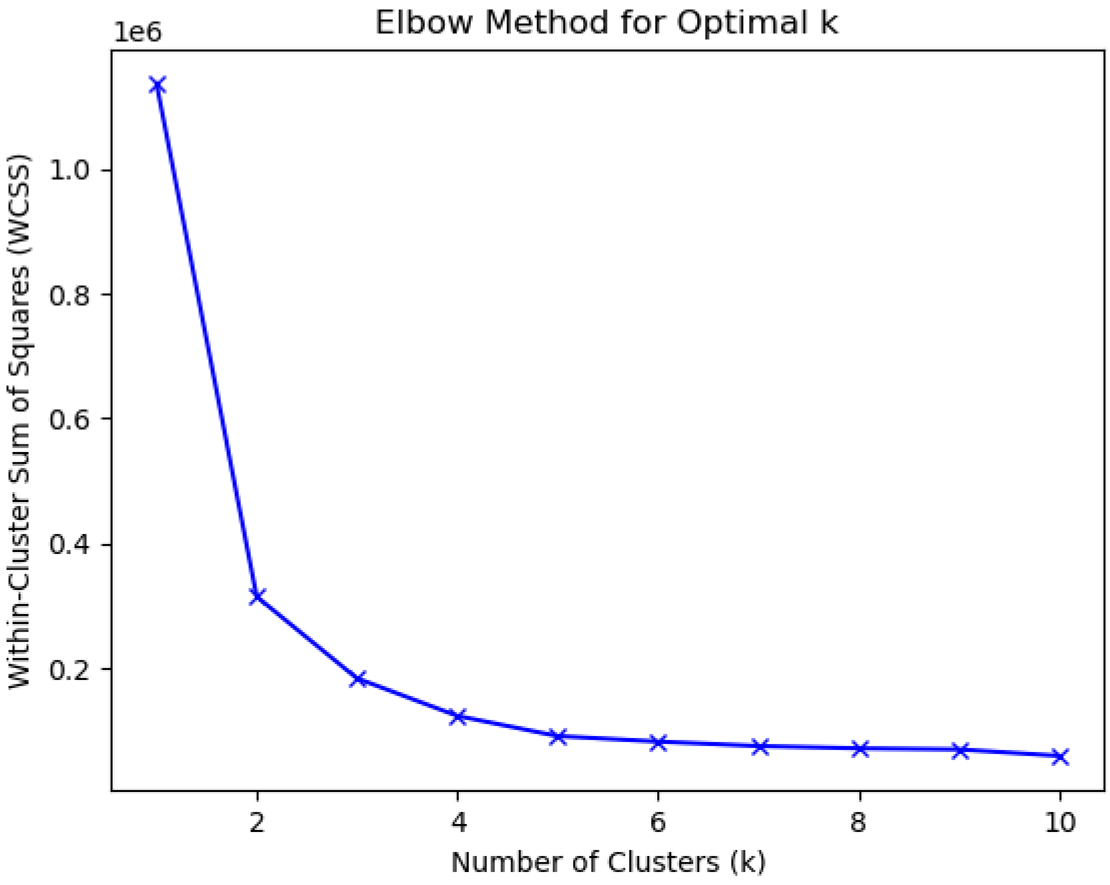
Results and Discussion

Quantitative Metrics and Visuals

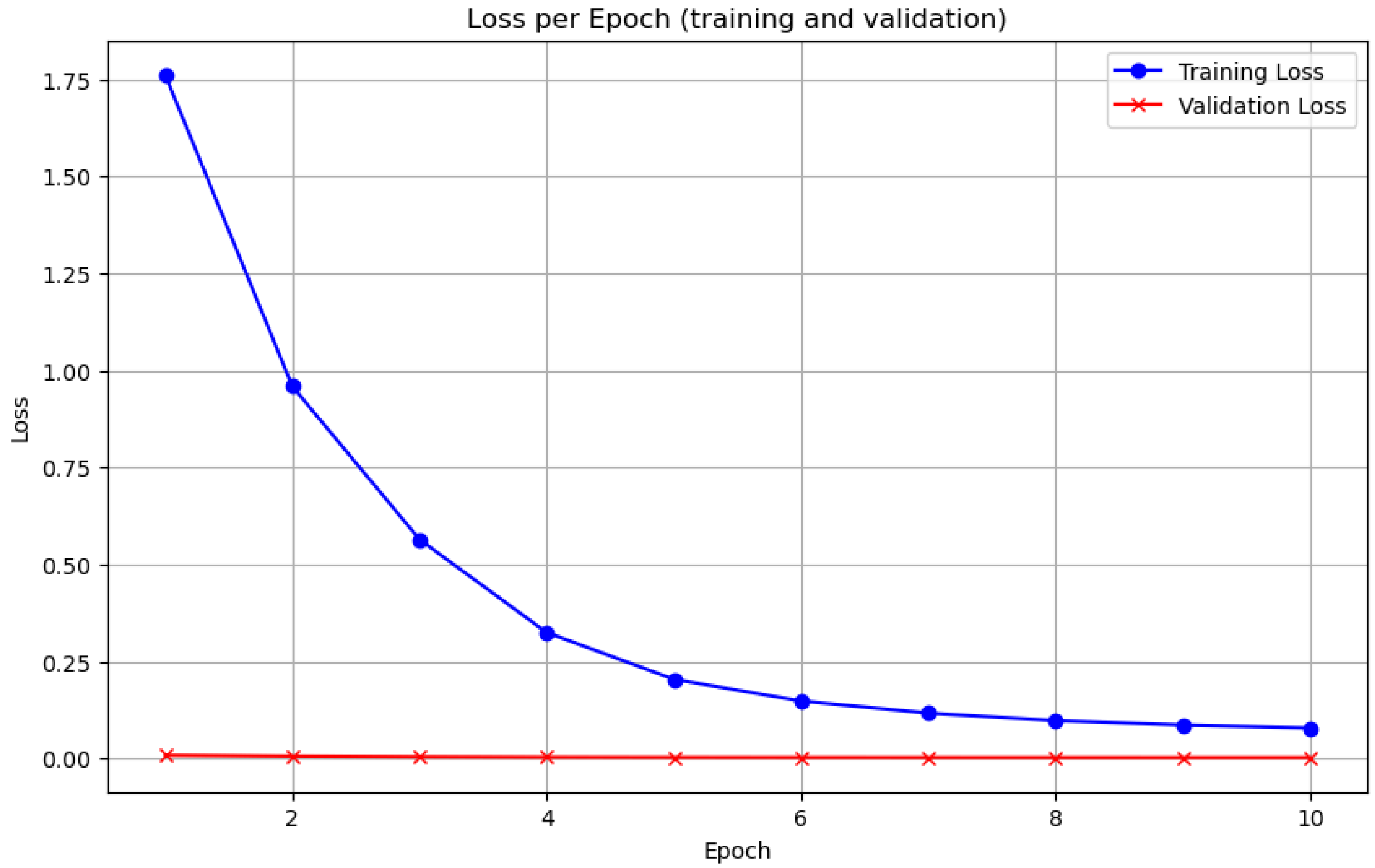
The performance of the models will be evaluated using metrics such as:

- K Means Elbow Plot: Computing the optimal numbers of clusters for our dataset
- Confusion Matrix: Visual representation of the correctness of the model.
- Mean Squared Error (MSE): Computing the mean of the square of the differences between the actual and predicted values when performing regression.
- Accuracy: Used to show the number of proper classifications when performing classification.

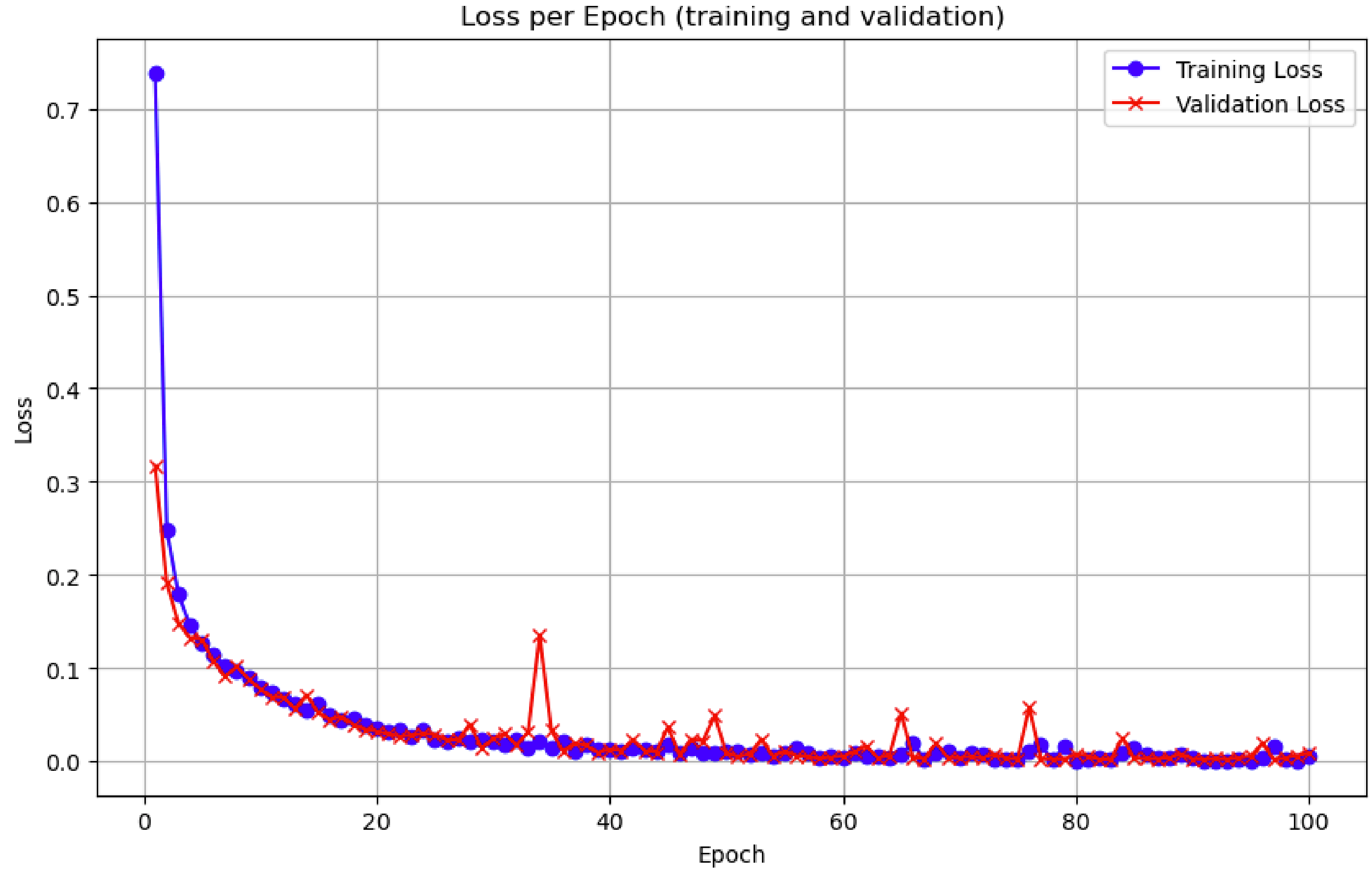
We were able to construct an elbow plot on our k-means clustering model to allow us to evaluate whether our assumptions about the ability to classify our data into at least some clusters was correct. This visual below at least indicated to us that it seems best to cluster our data into three groups as we see there is a visible decline in changing effectiveness after the 3 cluster mark.

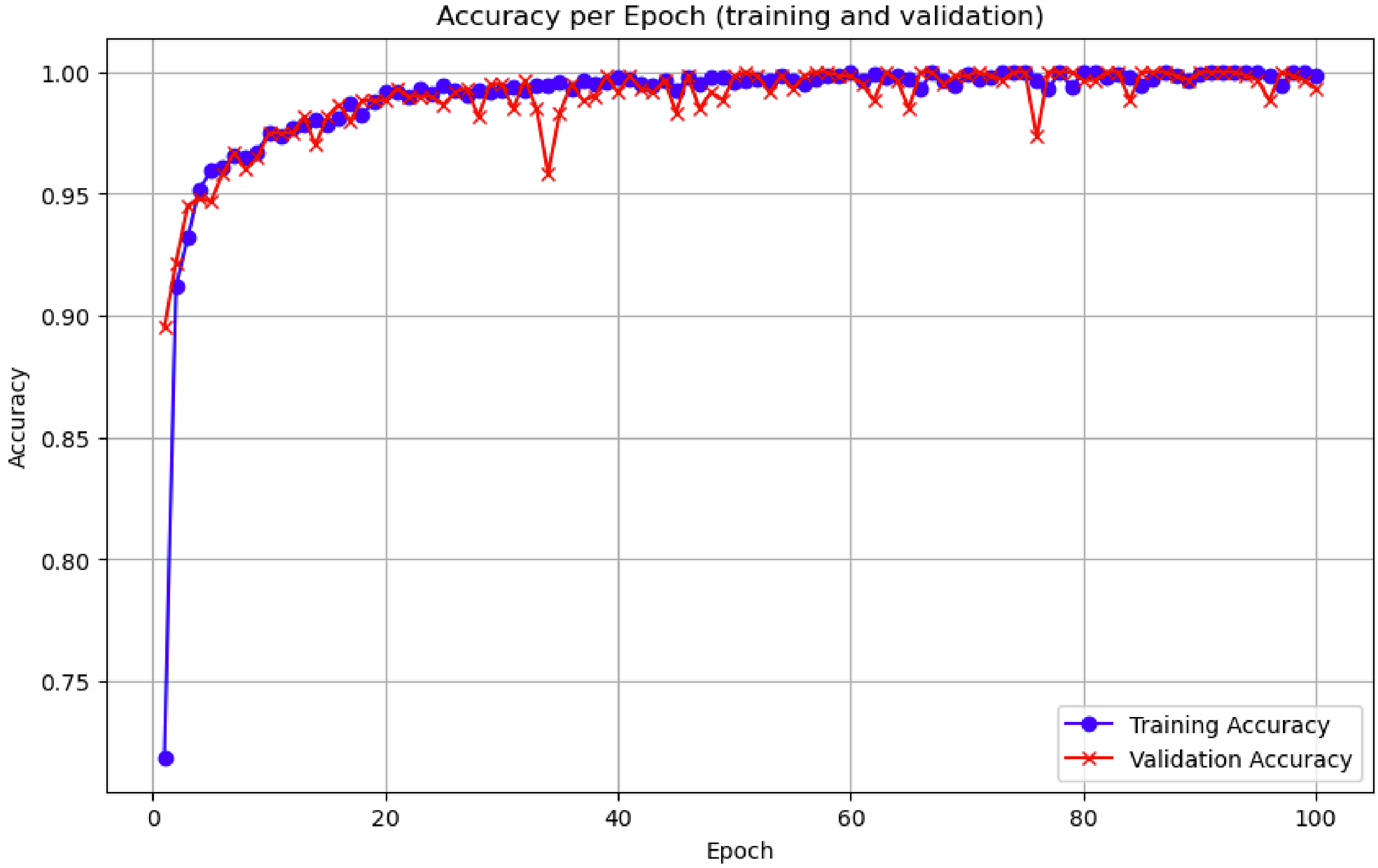


Additionally, we were able to get MSE values that decreased from 0.0083 to 0.0082. We can see that this value appears to fit our data well and through iterations performance improved. In the image below, we are able to see that there is a consistently low value for our validation loss and a consistent decrease in training loss which shows that the model the model appears to be effectively fitting the training data.

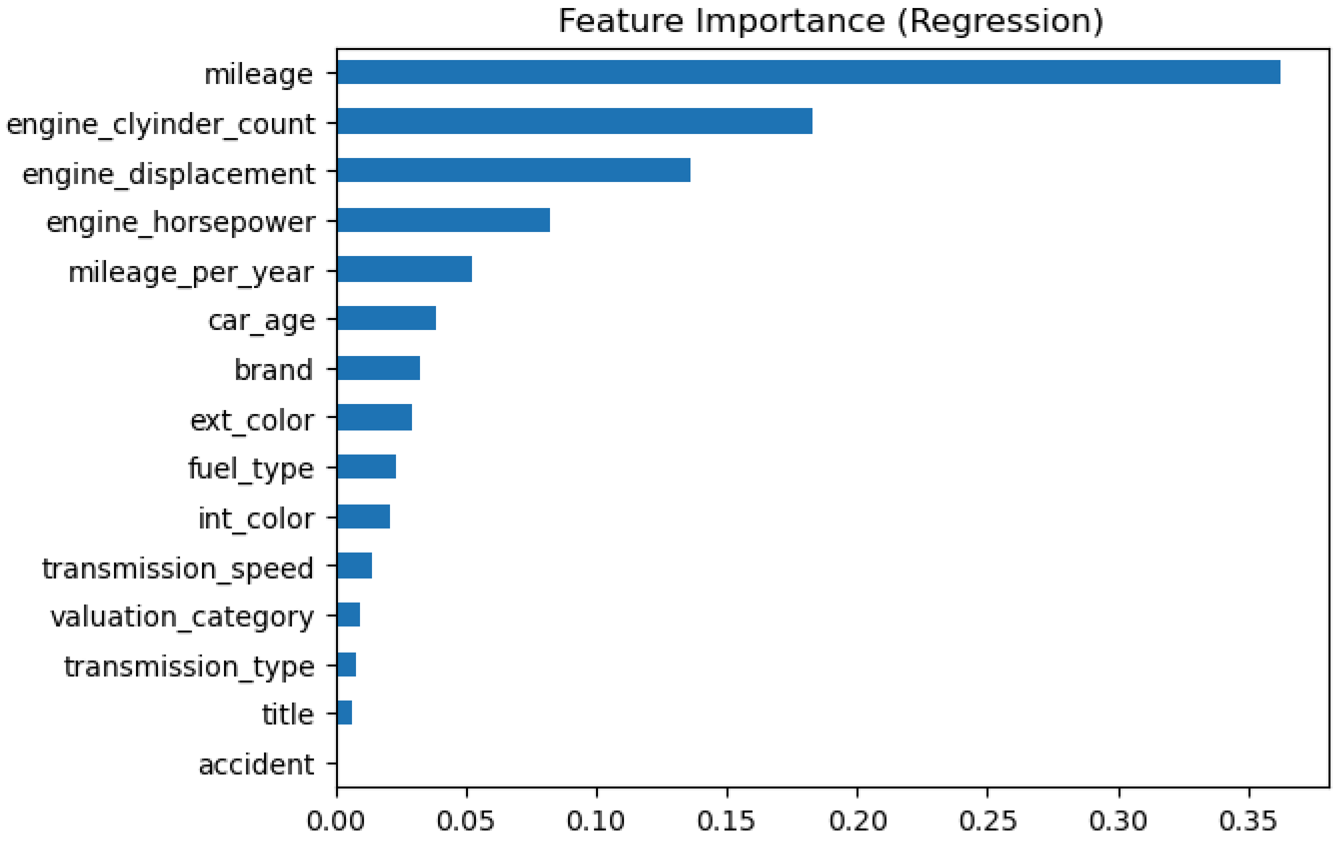


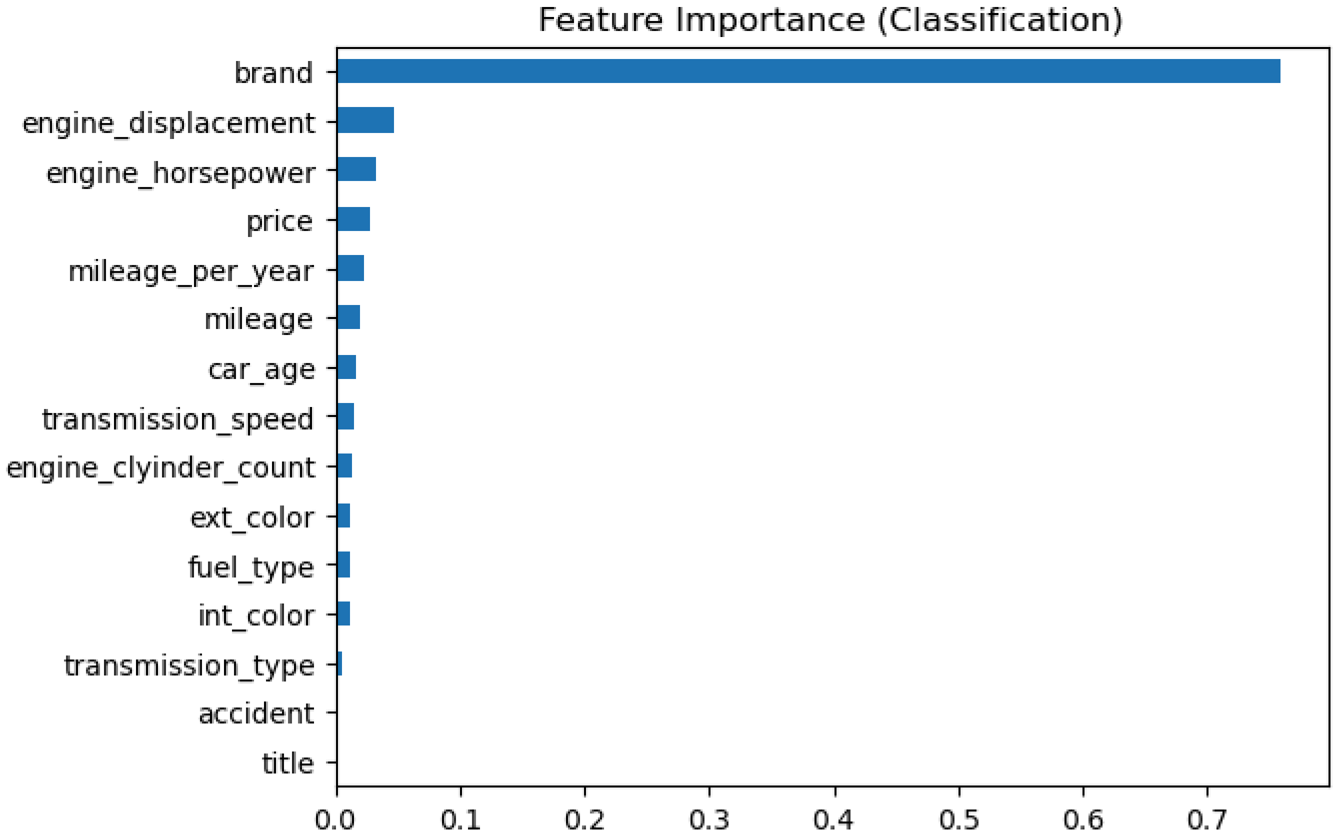
Next, when performing classification using the grouping formed by K-Means, were were able to achieve very high accuracy of '99.83%' on unseen testing data, and a test loss of 0.0038. We see that during our training and validation, we progress steadily over our epochs and eventually plataeu when reaching almost perfect accuracy. Overall, we were very happy with these results as they showed that categorizing our data when a viable strategy, and our model could learn very well from these categories. Below is an image of the training and validation loss graph and accuracy graph.



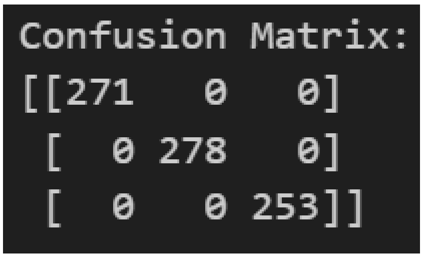


When conducting a random forest regression and classification, we wanted to be able to visualize the importance of different features to see how they may be related to the classification provided by our k-means analysis. The two graphs are below for regression and classification respectively.





We also wanted to construct a confusion matrix for the classification feature importance since we saw that it seemed a single feature was being used for classification. This image is below. Since all values fit along the diagonals, it does confirm -- along with the 100% accuracy -- that it is currently being classified solely by brand.



```
Confusion Matrix:
[[271   0   0]
 [  0 278   0]
 [  0   0 253]]
```

Analysis of Models and Challenges

K-means

To implement the unsupervised learning method, KMeans clustering, we first used the elbow method to determine the optimal number of clusters. Based on this method, we identified three clusters. After dividing the dataset into these three clusters, we applied different statistical methods to summarize each cluster. For categorical columns such as 'brand', 'fuel_type', 'transmission_type', and 'accident', we used frequency counts. For numerical columns such as 'car_age', 'mileage', 'mileage_per_year', 'engine_cylinder_count', 'engine_displacement', 'engine_horsepower', 'transmission_speed', and 'price', we calculated the mean, median, mode, minimum, maximum, skewness, and kurtosis. Based on these summaries, we classified the clusters as reasonably priced, moderately priced, and least reasonably priced. A summary of the classification from most reasonable to least reasonably priced, and the justification is provided below:

	Cluster	Description
0	Cluster 2	Lowest mileage, moderate car age, highest engine cylinder count, largest engine displacement. Higher-performance vehicles with less wear, justifying slightly higher prices if well-maintained.
1	Cluster 1	Older cars, lower cylinder counts, smaller engines, suggesting modest performance and lower value. Less intensive usage (mileage per year) and generally moderate mileage.
2	Cluster 0	Moderate engine displacement and age, higher mileage and yearly use, indicating more wear. Engine features don’t justify high-performance valuation.

We used these three clustering to label out data into valuation categories. Thus, our problem could be turned into a multi-class classificaiton setting.

Neural Networks

When we implemented two neural network models: one for regression to predict car prices, and another to predict vaulation category labels. Initially, we observed that the loss being measured was encountering numerical instability, causing it to transition to 'NaN' values. To resolve this problem, we had to revisit the data pre-processing steps to reduce the number of columns and our encoding strategies. For instance, we shifted from one-hot encoding to ordinal encoding and one-to-one mapping based encoding for discrete values columns such as car transmission and brand. Additionally, we added a min-max normalization scaler so that we can represent some columns within acceptable limits and not explode the loss to large values. After these changes, we observed that our loss, mean square error (when using regression), and cross entropy loss (when performing classication) values were within acceptable limits and were decreasing as iterations were progressing.

Random Forest

We utilized the random forest model for predicting car prices (regression) and classifying prices as reasonably priced, moderately priced, and least reasonably priced (classification).

For the random forest regressor model to predict car prices, we performed a grid search to estimate the best parameters for the model. We varied the number of estimators, which represent the number of trees in the random forest model, the height of the trees, the minimum number of samples required to split an internal node, and the minimum number of samples required to be a leaf node. Higher values for `min_samples_leaf` and `min_samples_split` reduce the model's complexity and potentially reduce overfitting. Increasing the number of trees improves the model's performance by reducing variance but also increases computational power and time. Deeper trees can lead to overfitting because they can model complex patterns in the data.

The random forest regression model achieved a **mean squared error of 2.6368** and an **R^2 value of 0.2010**. We also utilized the interpretability of this model to examine the feature importance scores. The features that contributed the most to the prediction values were mileage, engine cylinder count, engine displacement, engine horsepower, and engine power ratio. To summarize, the engine features and the mileage offered by a car were found to be strong determinants of the prices of used cars.

To implement the random forest classification model to classify car deals, we used SMOTE to handle class imbalance. We performed a similar grid search for the classification model. The feature importance values for the random forest model revealed that the brand of the car is a strong determinant in classifying the price of cars. This result aligns with our expectations because the brand of the car is strongly correlated with the engine of the car and the mileage offered by that engine. Therefore, the brand of the car plays an important role in determining whether the price of the model is fair or unfair for the end user.

Comparison of the Models

	Comparison Criteria	K-Means	Neural Network	Random Forest
0	Accuracy	Not applicable since the ground truth labels are unknown.	High accuracy because of its ability to capture relationships between attributes such as brand and model	High accuracy after adopting additional strategies for analysis such as SMOTE for class imbalance and grid

	Comparison Criteria	K-Means	Neural Network	Random Forest
				search for finding dominant features.
1	Learning Non-Linearly Separable	Spherical decision boundaries only	Non-linear relationships can be modeled.	Non-linear relationships can be modeled.
2	Training Time	Longer training time because it computes the distance between each pair of data points	Less because we were able to achieve convergence in a reasonable number of epochs with our architecture.	Long because we had to use a large number of estimators each having a sizeable depth.
3	Scalable	Not scalable to larger dataset due to the complexity of storing and calculating distances across each pair of data points. These need to fit into memory at each iteration.	Scalable as the algorithm processes data points in batches and learns through backpropagation to update weights. The memory required would be propotional to the batch size.	Not scalable to larger dataset because at each iteration we would have to find out the split that provides maximum information gain. This requires memory porportional to the number of data points.
4	Overfitting	We used to elbow method to ascertain the appropriate number of clusters, hence this does not overfit.	Hyperparameters can be tuned, such as regularization parameter, weight_decay, to reduce overfitting.	Hyperparameters available, such as depth and leaf criteria for minimum samples, to reduce overfitting as detailed in the analysis section.

Final Results & Potential Next Steps

Both the Neural Network and Random Forest performed extremly well on our data, with Random Forest slightly out perfoming the Nueral Network.

From the feature importatance and confusion matrix for the Random Forest model, we can see that in the classification setting, only the car brand played a role in the model's predictions. Upon inspecting the data, this was probably due to the wide array of car models with only three possible output labels. Car's with similar models were clustered together, allowing for the Random Forest to soley classify based on the model. On the other hand, we see that in the regression case, mileage and engine specs played a major role in the price estimation. The model relied on more features in this case to get a better understanding. From a consumer perspective, we can use this feature importance to guide prospective buyers to look into the enginge specs and consider if the features such as the number of cylinders and clyinder configuration are worth the additional cost.

In both cases, we saw that our classification setting achieved high accruacy, while in the regression task, there could be some improvements. Thus, some potential next steps might be additional feature engineering to get a better understanding of what features are introducing noise. From the Random Forest model, we gained insight that engine specs and mileage on the car played a large role. We could look into adding additional features to the data set relevent to the engine such as compression ratio, torque, etc. These might decrease loss and improve accuracy in the regression task. Moreover, we also saw the mileage was another important feature. We could engineer some new features such a miles per gallon based on the car model, make, and year to gain additional insights about the vehicle.

In summary, the Random Forest and Neural Network models demonstrated strong performance, with Random Forest excelling slightly, particularly in classification. The feature importance analysis highlighted the critical role of car brand in classification and engine specs in regression, providing valuable insights to assist consumer decision-making. While classification accuracy was high, regression performance still had some room for improvement through targeted feature engineering. Future work could explore incorporating additional engine-related features and creating derived metrics like miles per gallon to enhance predictive accuracy and reduce noise in the regression task.

References

[1] B. Hemendiran and P. N. Renjith, "Predicting the Prices of the Used Cars using Machine Learning for Resale," 2023 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS), Bhopal, India, 2023, pp. 1-5, doi: 10.1109/SCEECS57921.2023.10063133. keywords: {Machine learning algorithms;Pricing;Predictive models;Automobiles;Reliability;Task analysis;Regression tree analysis;Machine learning;forecast(predict);Random Forest;Decision Tree;Extra Tree Regressor;Bagging Regressor;Accuracy},

[2] J. Varshitha, K. Jahnavi and C. Lakshmi, "Prediction Of Used Car Prices Using Artificial Neural Networks And Machine Learning," 2022 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 2022, pp. 1-4, doi: 10.1109/ICCCI54379.2022.9740817. keywords: {Measurement;Machine learning algorithms;Linear regression;Artificial neural networks;Predictive models;Reliability theory;Prediction algorithms;ANN;keras;Used car price prediction;Regression;Random Forest;Machine Learning;Ridge;LASSO;Linear regression},

[3] Car Dealership Data Insights: Dataset Link

Gantt Chart

Gantt Chart link

Contribution Table

	Team Member	Proposal Contribution	Midterm Contribution	Final Checkpoint Contribution
0	Cesar Lopez Landaverde	Dataset collection & Description, Project goals, Streamlit App	Next Steps writeup, README writeup, combining of all diffrent scripts into one	Random Forest Implementation
1	Kailen Todd McCauley	Introduction & Background, Setting up GitHub repository, Video creation	Data Preprocessing, cleaning data, data loaders, updated data cleaning section in the writeup	Summarising results in report
2	Mohit Talreja	Unsupervised Learning, Supervised Learning, Streamlit App	supervised model coding, neural network implementation paragraph in writeup	Analysis of all the models
3	Trisha Jain	Problem definition & motivation, Project expectations, Streamlit App	unsupervised model coding, k-means paragraph implementation in writeup	Summarizing implementation of Random Forest in report
4	Uma Dukle	Data preprocessing, Results and metrics, Video creation & recording	visualizations, quantative measure in writeup	Visualisation and Final Video creation