COMS 4995 Applied Machine Learning

Group 4 - Project Proposal

Anqi Xia (ax2171), Azam Khan (ak4973), Uttam Gurram (ug2146)

Xiaokun Yu (xy2550), Yupei Shu (ys3597)

1. Background and Context to the problem statement

The ever-growing market of used cars is always an excellent place for individuals looking to purchase a car within their budget. So having a model that considers several car features to estimate the price point will be of valuable importance to the buyer. At the same time, for an individual looking to sell their car, it also provides the price point at which they can attract buyers' attention.

So our project's objective is to develop machine learning models to predict a car's price. Specifically, we aim to construct a regression model capable of estimating the listing price of a car. For this, we must carefully consider which features of a car are salient for accurate and reliable prediction of a car's price.

2. Dataset (US Used cars dataset)

The US Used Cars dataset is vast, containing roughly **3 million** records. The dataset includes information on cars listed since 2016. However, the value of cars depreciates over time, and at the same time, the actual listed price appreciates a bit because of inflation. So our model becomes unreliable if trained using the data from all the years, and we will limit our analysis to only the data of roughly **176K** records from 2021. The dataset contains 65 features, including numerical, categorical, and textual variables, providing a diverse range of information to analyze, and these features cover details about cars, sellers, and dealers.

The target variable is **price** to be predicted, and some of the features are listed below.

Numerical features	Categorical features	Textual features
back_legroom, bed_height, bed_length, city_fuel_economy, combine_fuel_economy, days_on_market, dealer_zip, etc	bed, body_type, cabin, truck, city, engine_type, exterior_color, engine_cylinders, fleet, etc	description: Car description on the car's listing page

3. Proposed ML techniques

Firstly, we need to perform data cleaning and fill in the missing data or drop the feature altogether if many of its values are missing. After that, we need to analyze the distribution of data and target variable with respect to the features. Also, we need to check for highly correlated features and drop one from each pair. We can also reduce the number of features using dimensionality reduction techniques like PCA. We can also use some state-of-the-art NLP models to perform sentiment analysis on the textual feature and convert it to a categorical feature. In the end, we can encode the categorical features into numerical features.

We plan to train and use **Linear Regression** model as a baseline. We will also train and evaluate other regression models such as **Lasso**, **Ridge**, and **Elastic-Net**. **Decision trees** will be trained and improved using Bagging and Gradient Boosting techniques. We will also train a simple neural network to compare the deep learning model performance with other models. Metrics such as AuROC score and RMSE will be used to compare the performance of different models. Also, we will implement hyperparameter tuning using Grid and Random search to obtain the best-performing models.