Emil Khuu

DESCRIBE THE PROBLEM

SCOPE

The goal of this project is to develop a ML based software product that estimates the value of a used car. The intention is to integrate the final solution into a web application, allowing users to input various car attributes and then receive an estimated price predicted by the ML model. Without machine learning, car pricing is often done by manually checking similar car listings . The performance will essentially be measured by the ML model's accuracy of price estimates. A tentative timeline of this project will be to first do data collection and preprocessing, then select multiple algorithms for training to find the best fit for our use case and data by using metrics like RMSE R^2, and MAE. Lastly, the model will be deployed to the platform and integrated with the user interface. I plan to train the model on Kaggle then deploy it using Gradio.

METRICS

Price estimates must be within 20-30% of the actual market value of the car to be considered a success. R^2, RMSE and MAE will be used as metrics to calculate the accuracy of the model. These metrics can be calibrated to stay within the acceptable error margins for business use.

DATA

The dataset will contain labels such as brand, model, mileage, year to determine a car's depreciation and market value, fuel type, engine size, transmission, accident history & clean title for indicators of a car's resale value. The data is sourced from a Kaggle dataset containing information on used car listings and since the data is publicly available on Kaggle, there are no significant privacy concerns. There were three labels that had missing values, which were imputed. All categorical variables (brand, fuel type, transmission, exterior color, interior color, accident & clean title) were encoded using label encoding. Continuous variables (mileage, year & price) were scaled to improve model performance.

MODELING

I plan to explore linear regression, random forest regression & gradient boosting. Initial performance will be measured using linear regression to estimate baseline R^2 and RMSE. This will be used to compare with the other two models. The baseline scores will be used for optimization purposes, for instance by going back to preprocessing.

DEPLOYMENT

The model will be deployed with Gradio. Users will input car attributes, and the model will return an estimated price based on predictions. After deployment, the model's performance will be continuously monitored for accuracy. If the predictions deviate significantly from expected values, the model will be retrained.