DECISION 520Q.101D

Team 29

Jojo Dong; Ayush Kumar; Yolanda Li; Colette Ma; Syed Irtiza Mehdi

Predicting Selling Price of Used Vehicles

Business Problem

Our business problem was focused on creating a value-determining machine learning model which would accurately predict the selling price of a particular used vehicle, given key characteristics, such as the make, model, mileage, age (year), fuel type, and the transmission being used. By being able to predict the price a car holds within the market, we aim to help dealerships, individual sellers, and online marketplaces in pricing their vehicles optimally, so as to maximize their profitability, alongside remaining competitive within the market. Additionally, the predicted price serves as a benchmark to what the car can be sold at, and when a certain party is looking to buy a certain car to flip it, especially in the case of dealerships, they will look towards purchasing the vehicle at a price lower than what the model predicts. Hence, not only does our model enable stakeholders to avoid the risks of underpricing, which would lead to revenue loss, or pricing the vehicle at a higher price, which would also drive the customers away, it also provides them an idea of how much to buy the car for, to be able to make a profit on the transaction.

Accordingly, the used car market generates a massive amount of data, where the traditional pricing models, such as simple averages or formulas, are unable to capture the complex relationships between the different features of the vehicle and the prices in the market. The inherent confusion within these markets relating to used/second-hand vehicles makes it difficult to determine a particular price where the seller is able to secure its competitive position.

Characteristics, like the mileage, make or age of a car interact differently with one another and influence the price of a vehicle in non-linear ways that the old models are not able to account for. However, machine learning models are able to work through large data sets, determine varying relationships between different variables, and handle nonlinear relationships to provide a more effective method of predicting prices. By leveraging such models, we can benefit all involved stakeholders within the automotive market, by pricing their vehicles optimally and responsibly, so that the factor of asymmetric information is also minimized, and all the parties are able to make a much more well-informed decision.

Data Understanding

a. Overview of the Dataset

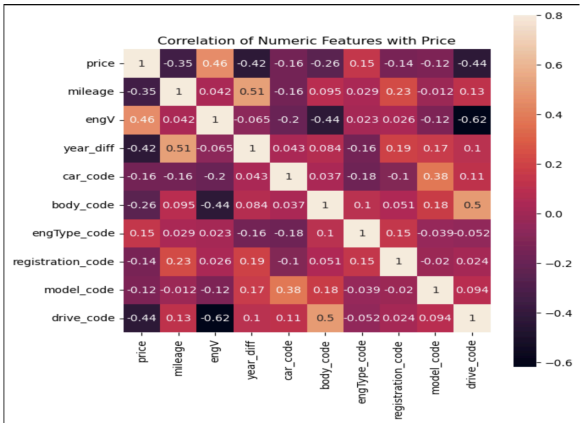
The dataset contains 8,739 records, each representing a car advertisement. It includes 10 variables related to various attributes of the cars being sold. The dataset contains several key variables: **car** (the brand or make), **price** (the asking price in USD), **body** (car body type such as sedan, crossover, or van), **mileage** (the number of kilometers or miles driven), **engV** (engine volume in liters), and **engType** (the type of fuel like Gas, Petrol, or Diesel). Additionally, it includes **registration** (indicating if the car is registered), **year** (manufacture year), **model** (specific car model), and **drive** (the drivetrain type, such as full, rear, or front).

b. Data Quality Considerations

Missing Values and data types: There are no missing values in this dataset based on the initial inspection. The data types seem appropriate for all columns—price, mileage, and engV are numeric, while car, body, engType, registration, model, and drive are categorical.

Data Preparation

- Data Encoding : We factorized (encoded) the categorical variables into numeric codes, and computed the difference between the current year and the car's manufacturing year. Rather than using the raw year, the age of the car becomes a more intuitive feature.
- Feature Scaling : Scaling features like mileage and engine volume is essential because their ranges may differ significantly.
- Outlier Detection and Removal : We identified and removed outliers using the interquartile range (IQR) method.
- Feature Correlation Analysis : We analyze the correlation between numeric features and the target variable (price). By understanding which features have a strong relationship with price, we can focus on those features to improve model accuracy.



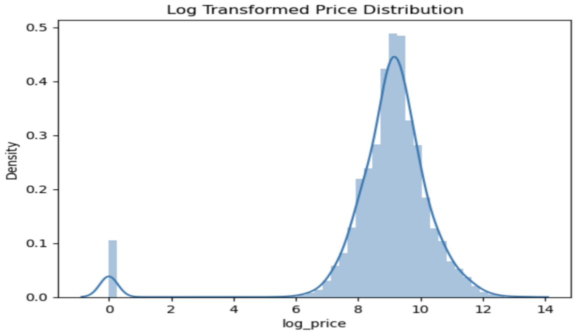
	price	mileage	engV	year_diff	car_code	body_code	engType_code	registration_code	model_code	drive_code
price	1	-0.35	0.46	-0.42	-0.16	-0.26	0.15	-0.14	-0.12	-0.44
mileage	-0.35	1	0.042	0.51	-0.16	0.095	0.029	0.23	-0.012	0.13
engV	0.46	0.042	1	-0.065	-0.2	-0.44	0.023	0.026	-0.12	-0.62
year_diff	-0.42	0.51	-0.065	1	0.043	0.084	-0.16	0.19	0.17	0.1
car_code	-0.16	-0.16	-0.2	0.043	1	0.037	-0.18	-0.1	0.38	0.11
body_code	-0.26	0.095	-0.44	0.084	0.037	1	0.1	0.051	0.18	0.5
engType_code	0.15	0.029	0.023	-0.16	-0.18	0.1	1	0.15	-0.039	-0.052
registration_code	-0.14	0.23	0.026	0.19	-0.1	0.051	0.15	1	-0.02	0.024
model_code	-0.12	-0.012	-0.12	0.17	0.38	0.18	-0.039	-0.02	1	0.094
drive_code	-0.44	0.13	-0.62	0.1	0.11	0.5	-0.052	0.024	0.094	1

- Normalization and Johnson SU Distribution Fitting : We normalized the data, then we found that the Johnson SU Distribution fits the data best, and it was the best way to reduce skewness.
- Feature Engineering : We generated interaction terms such as mileage * year_diff and engV * year_diff. These interaction terms can capture more complex relationships between features and enrich our database.
- Split the Dataset into Train and Test Sets.

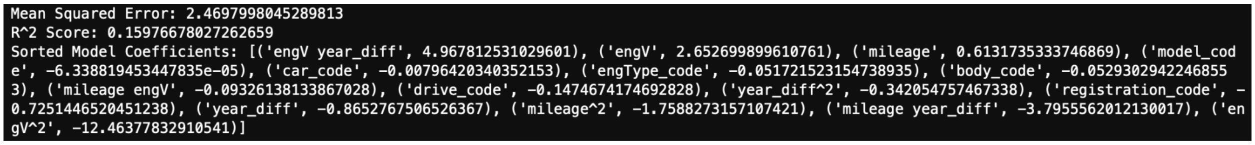
Modeling

1. Linear Regression

- Data Preprocessing and Feature Engineering: We applied a logarithmic transformation to the target variable (price) to reduce skewness and normalize the data. After scaling continuous features with MinMaxScaler, we generated polynomial features of degree 2 for mileage, engV, and year_diff. We fit the model with linear regression, finding the relationship between the features and the log-transformed price.



- Model Evaluation: R² Score: The model's R² score is 0.1597, meaning the model explains about 15.97% of the variance in the data. This indicates that the model is not capturing all the complexity in the data and could benefit from further refinement.



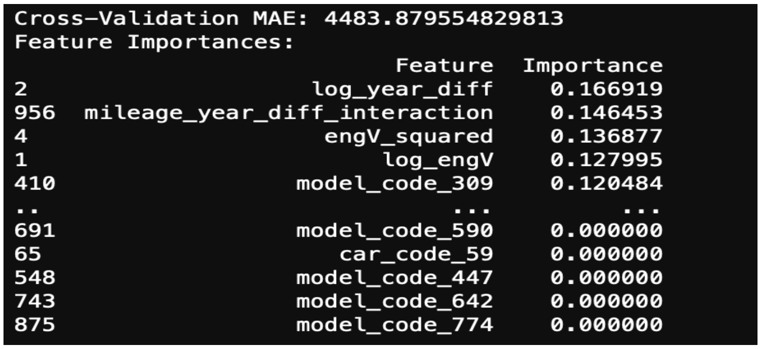
```
Mean Squared Error: 2.4697998045289813
R^2 Score: 0.15976678027262659
Sorted Model Coefficients: [('engV year_diff', 4.967812531029601), ('engV', 2.652699899610761), ('mileage', 0.6131735333746869), ('model_cod
e', -6.338819453447835e-05), ('car_code', -0.00796420340352153), ('engType_code', -0.051721523154738935), ('body_code', -0.0529302942246855
3), ('mileage engV', -0.09326138133867028), ('drive_code', -0.1474674174692828), ('year_diff^2', -0.342054757467338), ('registration_code', -
0.7251446520451238), ('year_diff', -0.8652767506526367), ('mileage^2', -1.7588273157107421), ('mileage year_diff', -3.7955562012130017), ('en
gV^2', -12.46377832910541)]
```

- Mean Squared Error (MSE): The MSE is 2.4698, which indicates the average squared difference between the predicted and actual values. While the model is functional, its error rate suggests that more sophisticated techniques may be needed.
- Linear Regression with Polynomial Features: For pros, Linear regression provides a straightforward interpretation of feature importance through its coefficients. But as seen from the relatively low R² score, the model may be underfitting the data.

- Why This Model: Linear Regression is a solid starting point because it provides interpretability, which is crucial when making data-driven business decisions. Stakeholders can easily understand which features drive car prices up or down.

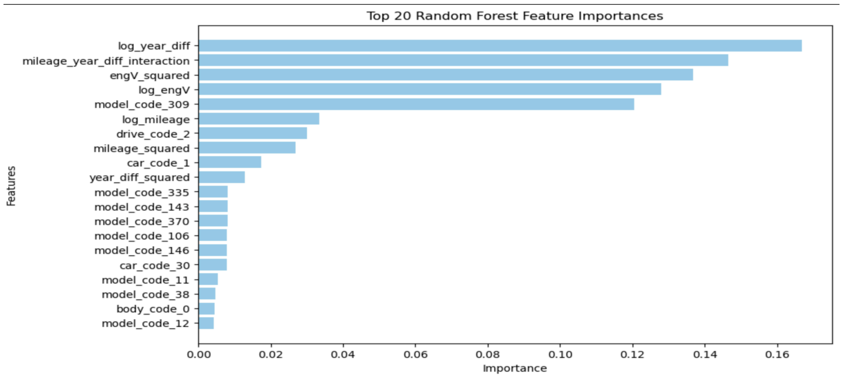
2. **Random Forest Regression**

- Feature Engineering: We applied logarithmic transformations to continuous features (mileage, engV, year_diff) to reduce skewness, while squared terms captured non-linear relationships. Also we created interaction between mileage and year_diff to capture potential combined effects. For the categorical features such as car_code, body_code, and model_code, we applied one-hot encoding to fit the model.We used a Random Forest model with 100 decision trees (n_estimators=100) to capture non-linear relationships between the features and the target variable (price).
- Model Evaluation: The Random Forest model resulted in an MAE of **4483.88**, which is an improvement over simpler models like Linear Regression. This indicates the model's predictions are, on average, off by approximately 4483 units from the actual car price.



```
Cross-Validation MAE: 4483.879554829813
Feature Importances:
                           Feature  Importance
2                     log_year_diff    0.166919
956  mileage_year_diff_interaction    0.146453
4                      engV_squared    0.136877
1                          log_engV    0.127995
410                  model_code_309    0.120484
..                              ...         ...
691                  model_code_590    0.000000
65                      car_code_59    0.000000
548                  model_code_447    0.000000
743                  model_code_642    0.000000
875                  model_code_774    0.000000
```

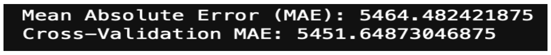
The top contributing features include log_year_diff (age of the car), mileage_year_diff_interaction, and engV_squared, which suggests that vehicle age, mileage, and engine volume are the most significant predictors of car price.



- Why This Model: Random Forest effectively handles the non-linear relationships and high-dimensional data in your dataset. The model's flexibility and robustness to outliers make it well-suited to predicting car prices, which are influenced by numerous factors.

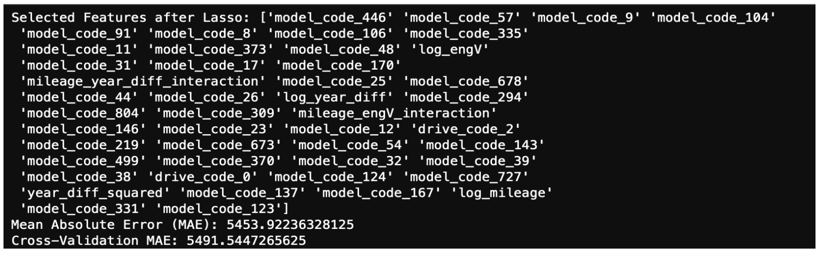
3. XGBoost

- Feature Engineering: We applied logarithmic transformations to continuous features while squared terms captured non-linear relationships. Also, we applied Lasso regularization to select important features, reducing dimensionality. We used XGBoost with 100 trees (n_estimators=100), a learning rate of 0.1, and a maximum depth of 6 for each tree.
- Model Evaluation:The model achieved an MAE of **5464.48**, indicating that the average error in predicting car prices is around 5464 units. This is an acceptable result considering the complexity of the dataset.



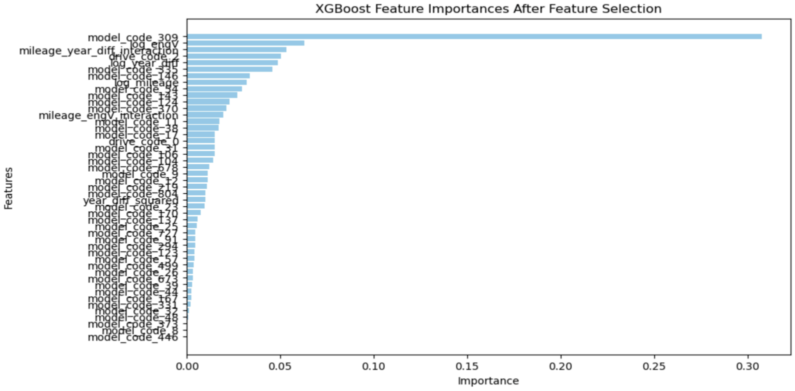
```
Mean Absolute Error (MAE): 5464.482421875
Cross-Validation MAE: 5451.64873046875
```

The cross-validation score was slightly lower at **5451.64**, confirming that the model is generalizing well to unseen data.



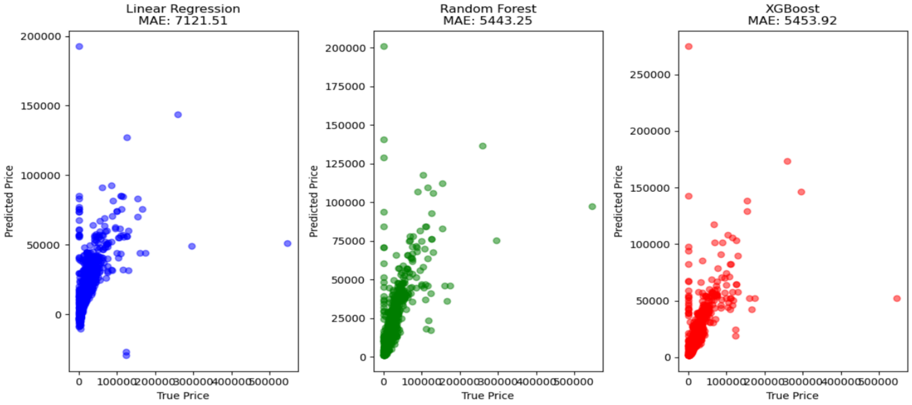
```
Selected Features after Lasso: ['model_code_446' 'model_code_57' 'model_code_9' 'model_code_104'
 'model_code_91' 'model_code_8' 'model_code_106' 'model_code_335'
 'model_code_11' 'model_code_373' 'model_code_48' 'log_engV'
 'model_code_31' 'model_code_17' 'model_code_170'
 'mileage_year_diff_interaction' 'model_code_25' 'model_code_678'
 'model_code_44' 'model_code_26' 'log_year_diff' 'model_code_294'
 'model_code_804' 'model_code_309' 'mileage_engV_interaction'
 'model_code_146' 'model_code_23' 'model_code_12' 'drive_code_2'
 'model_code_219' 'model_code_673' 'model_code_54' 'model_code_143'
 'model_code_499' 'model_code_370' 'model_code_32' 'model_code_39'
 'model_code_38' 'drive_code_0' 'model_code_124' 'model_code_727'
 'year_diff_squared' 'model_code_137' 'model_code_167' 'log_mileage'
 'model_code_331' 'model_code_123']
Mean Absolute Error (MAE): 5453.92236328125
Cross-Validation MAE: 5491.5447265625
```

XGBoost provides a feature importance ranking, with features such as model_code_309, mileage_year_diff_interaction, log_engV being the most influential in predicting car prices.



- ○ Why the model: Given the large number of categorical variables (e.g., model_code, drive_code) and interaction terms, XGBoost's ability to handle high-dimensional data ensures that the model makes accurate predictions while considering many variables. With the use of Lasso (L1) regularization and the gradient boosting approach, XGBoost minimizes overfitting, leading to better generalization to unseen data.

Evaluation



The three models—Linear Regression, Random Forest, and XGBoost—have been evaluated using MAE and visualized by plotting their predicted prices against true prices.

For linear Regression, the MAE of 7121.51 demonstrates that linear regression struggles to capture the complex relationships in the dataset, evident from the spread of the predicted values in the plot. With an MAE of 5443.25, the Random Forest model performs significantly better than Linear Regression, capturing the non-linear relationships in the data more effectively. The MAE of 5453.92 is very close to that of Random Forest, and visually, the predictions are similarly concentrated around the true prices. While XGBoost is generally expected to outperform Random Forest, in this case, it shows similar performance. This may be due to the specific data characteristics or further tuning being required.

By minimizing the error between predicted and true prices, the business can reduce overpricing and underpricing, while offering fair and market-driven prices to enhance customer experience. In terms of ROI, the business could calculate the improvement in revenue generated by reducing pricing errors (e.g., increasing sales volume by X% through optimized pricing). The cost of developing and maintaining these models includes the initial data preparation, model training, deployment, and ongoing maintenance.

Deployment

One of the ways the prediction model can be deployed is through integrating it into a web-based platform where the users are able to manually input key vehicle characteristics, such as the mileage, model, year, and engine size and receive an instantaneous price estimate of their vehicles, or while browsing through the listings, there could be a section indicating the ‘estimated fair value’ predicted through the model. Accordingly, this particular model can also be effectively utilized in e-commerce platforms, where the sellers are able to use this model in order to list their vehicles at accurate, data-driven prices and for buyers to gather an idea of the fair market value of the vehicle, increasing transparency and trust, and removing that barrier of asymmetric information by providing both sides an estimate of the vehicle's value. Dealerships can also effectively deploy this model in setting competitive prices for their present inventory, as well as for trade-in evaluations, making the transaction more consistent and enhancing the customer experience. It can also be used as a membership/subscription-based application in mobile phones, allowing users to get pricing estimates on the go and make informed purchasing/selling decisions.

While trying to deploy our prediction model into an existing web-based platform, we should ensure that the model is able to seamlessly handle large volumes of inputs and continue to accurately predict the pricing estimates. There will be certain instances that the model has not been trained on, and there should be methods of tackling such challenges with the human element involved. Technical challenges, relating to scaling the model and being able to handle multiple users simultaneously, especially in high-traffic periods, may also arise if the model is not deployed effectively. It is essential that the model is able to be integrated into various platforms and devices, and utilized in multiple use cases. It is highly important that the firm gets the model updated regularly, which includes retraining the model with updated data to continue producing accurate pricing, effectively adjusting to the dynamic nature of the automotive industry, given the shifting trends, such as changes in demand for petrol vehicles against electric vehicles, and the sentiment within the general market due to the fluctuating economic environment.

One of the major concerns in terms of the ethical considerations is definitely related to the data privacy issues, where the vehicle sellers may be reluctant in sharing private vehicle information (e.g., registration history), especially if they are unsure about how the data will be stored and/or used. It is critical that the owners are made aware of the process of how the data is collected, processed and secured in order to build that trust. The users' personal information must be protected through implementation of robust encryption techniques, and the firms collecting such information should fully comply with the data protection regulations, ensuring that the user content is obtained, and that the data is not sourced to other firms, but only used for the intended purpose of price estimation.

It is important to consider the inherent nature of the dynamic automotive industry while looking to deploy the model, especially considering the rise in Elective Vehicles (EVs) and other such innovations. The data being provided to the model should be updated frequently, as surges in demand for EVs or other environmentally friendly alternatives could greatly impact the general pricing of vehicles within the market, and have a negative impact on the model's prediction accuracy. To ensure that the model remains effective, continuous monitoring of the model's performance is required, which may also be done through taking user feedback of prices in their deals and transactions, and continuously training the model given those values once the data is verified. This ensures that the model is able to account for current market trends, emerging technologies, and consumer behavior while looking to predict prices. Additionally, firms may also consider implementing a system where the users are able to report inaccurate estimates, allowing for real-time adjustments and model's fine-tuning.

Appendix

Yolanda Li - modeling, evaluation, project proposal

Ayush Kumar - status report, presentation deck

Syed Irtiza Mehdi - business problem, deployment

Colette Ma - data preparation, status report, presentation deck

Jojo Dong - presentation