# END-TO-END MACHINE LEARNING PROJECT

(Predicting used car prices with combining multiple boosting algorithms)

(Prepared by: Emir Tatlıcı)

In the rapidly developing automotive market, accurate car price prediction plays a crucial role in guiding purchase and sales decisions for various stakeholders. This project presents an end-to-end machine learning regression model developed to predict car prices based on a comprehensive set of influencing factors, including engine size, mileage, brand, and transmission type. By leveraging a rich dataset of historical car sales, the model aims to unveil the relationships between these attributes and their impact on vehicle pricing. The performance of the model is evaluated using Root Mean Squared Error (RMSE) as the primary metric, which quantifies the average prediction error and provides insights into the accuracy of the model's forecasts. The results demonstrate the model's efficacy in offering reliable price predictions, thereby empowering dealerships and potential buyers to make informed decisions.

Evaluation Metric

$$RMSE = sqrt\left(\left(\frac{1}{n}\right) * \Sigma(y_i - \hat{y}_i)^2\right)$$

**Steps:**

* Data Imports

*Train-Validation-Test Split

* Feature Engineering

* Data Preprocessing

* Pipeline Build for ML algorithms

* Hyperparameter Optimization (Optuna)

* K-Fold Cross-Validation

* Model Evaluation

* Weight Optimization

* Combining Models

* Conclusion & Insights

**Data Characteristics**

The training set consists of 203,034 rows and 12 columns, while the test set has 50,759 rows and 12 columns. The columns in the dataset include publishDate, firstPublishedDate, state_code, zip, price, mileage, make, body_type, vehicle_style, vehicle_trim, year, and listed_days_count. The training set is divided into 80% for training and 20% for validation. The test set is used only for calculating metrics and not for model fitting.

Feature engineering has been applied to enhance the dataset, resulting in additional columns: engine_size, door_count, cylinder_count, transmission_type, car_age, transmission_type_2, and age_segment. publishDate and firstPublishedDate columns are not used throughout the whole process.

**Hyperparameter Optimization:** To identify the best hyperparameters within specified ranges, I utilized the open-source Optuna Python library. Optuna employs Bayesian Optimization to effectively search for optimal hyperparameters within a defined hyperparameter space.

I used 2 ensemble algorithms for their Respective Speed and High Accuracy. LigthGBM and XGBoost

**XGBoost:** Best RMSE Score: 8117.35, hyperparameters:

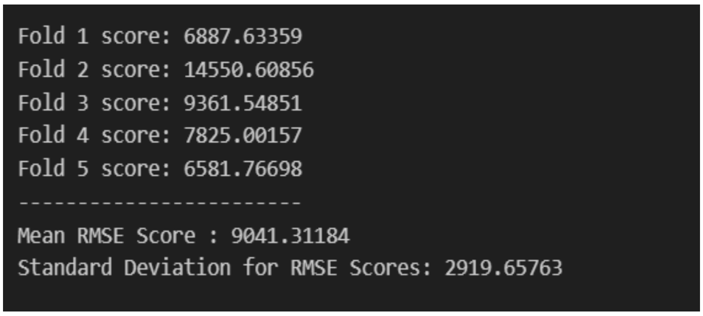
{'max_depth': 15, 'learning_rate': 0.05296660283065631, 'n_estimators': 126, 'min_child_weight': 10, 'subsample': 0.9893631526916168, 'colsample_bytree': 0.5124428472442656, 'gamma': 1.316880000486226e-08, 'reg_alpha': 0.5298860553217419, 'reg_lambda': 4.136887971660181e-06}

**LightGBM:** Best RMSE score: 7886.35, hyperparameters:

{'max_depth': 10, 'learning_rate': 0.13041244380537204, 'n_estimators': 598, 'min_child_weight': 7, 'subsample': 0.6557583384866239, 'colsample_bytree': 0.5917814788649068, 'reg_alpha': 1.155243040648194e-08, 'reg_lambda': 0.03998522636161519, 'num_leaves': 137, 'min_child_samples': 40, 'feature_fraction': 0.5020194969680079}

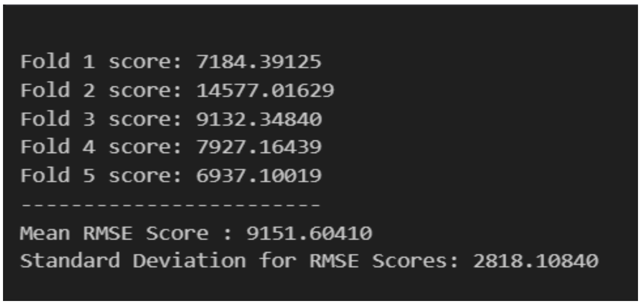
**Evaluating Models with K-Fold Cross Validation**

XGBOOST: 5-Fold Cross Validation Scores



```
Fold 1 score: 6887.63359
Fold 2 score: 14550.60856
Fold 3 score: 9361.54851
Fold 4 score: 7825.00157
Fold 5 score: 6581.76698
------------------------
Mean RMSE Score : 9041.31184
Standard Deviation for RMSE Scores: 2919.65763
```

LightGBM: 5-Fold Cross Validation Scores



```
Fold 1 score: 7184.39125
Fold 2 score: 14577.01629
Fold 3 score: 9132.34840
Fold 4 score: 7927.16439
Fold 5 score: 6937.10019
------------------------
Mean RMSE Score : 9151.60410
Standard Deviation for RMSE Scores: 2818.10840
```

Both models have high RMSE scores in Second Fold which suggests that the models we built were struggling to predict the car prices on that Fold samples. This might occur because of the high priced cars because small percentage of differences in high prices could lead to higher RMSE Scores in that respective Fold.

After looking at both Cross Validation results I can say that both models work very similarly. LightGBM outperforms XGBoost with small RMSE scores. In the next step I will be applying weight optimization in validation set for better model with those weights I will build combined model which will use both models in itself.

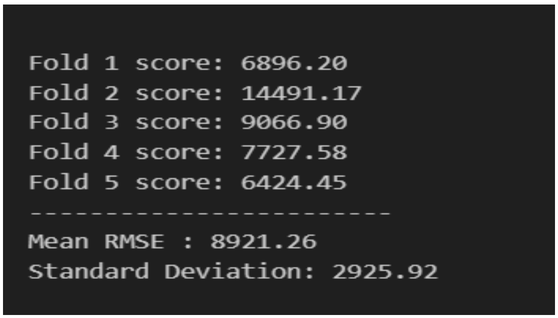
After applying weight optimization to models for better score Optuna suggested me these weights for the models

w1 0.639 (LightGBM)

w2 0.361 (XGBoost)

After combining the weights of each model I built another model and Applied 5-Fold cross validation to that.

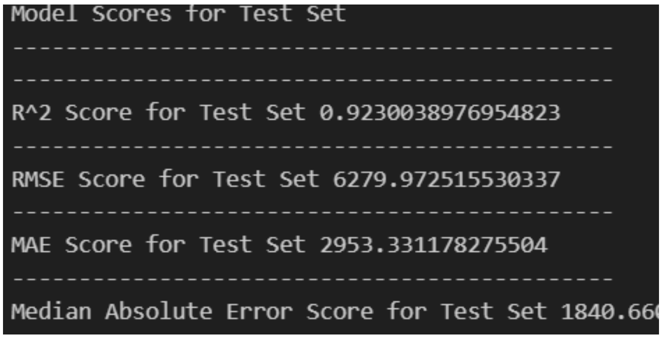
Combined_Model 5-Fold Cross Validation Scores



```
Fold 1 score: 6896.20
Fold 2 score: 14491.17
Fold 3 score: 9066.90
Fold 4 score: 7727.58
Fold 5 score: 6424.45
------------------------
Mean RMSE : 8921.26
Standard Deviation: 2925.92
```

After combining both models with the suggested weights, we successfully outperformed the individual model results on the validation set. Now, I plan to combine both the training and validation sets to enhance model convergence for the test set(totally unseen data).

Test Data Results!



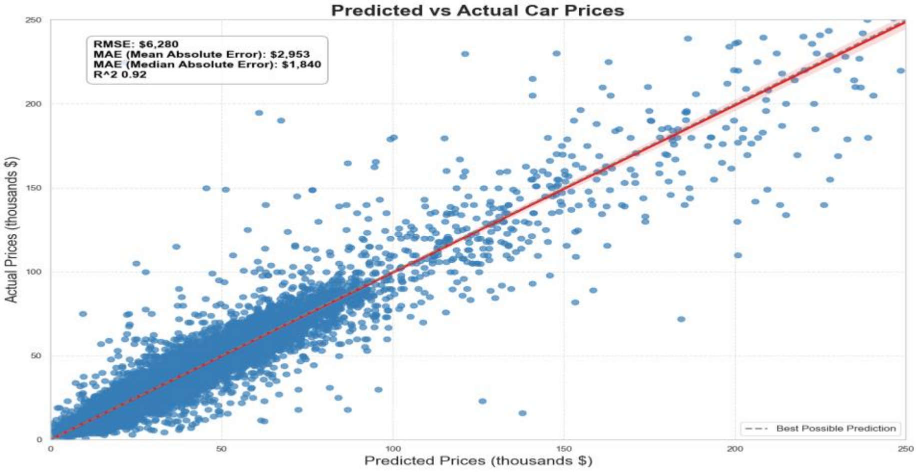
```
Model Scores for Test Set
-----------------------------------------
-----------------------------------------
R^2 Score for Test Set 0.9230038976954823
-----------------------------------------
RMSE Score for Test Set 6279.972515530337
-----------------------------------------
MAE Score for Test Set 2953.331178275504
-----------------------------------------
Median Absolute Error Score for Test Set 1840.66
```

We successfully reduced our RMSE score, surpassing the results achieved on the validation set. This improvement indicates our model's enhanced predictive accuracy. Additionally, the R² score of 0.92 suggests that our models can explain 92% of the variability in car prices based on the input features.

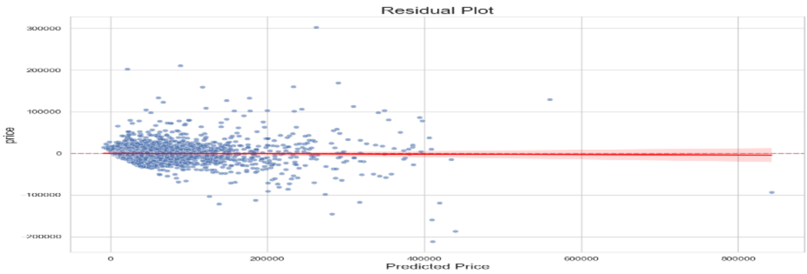
The Mean Absolute Error (MAE) of 2953 signifies that, on average, our model's predictions deviate from the actual car prices by about $2953. Meanwhile, the Median Absolute Error (MedAE) of 1840 indicates that half of the predictions are within $1840 of the actual prices, reflecting a more robust performance against outliers in the dataset.

## Deep Diving in to Test Results

Visualization of predicted and actual prices.



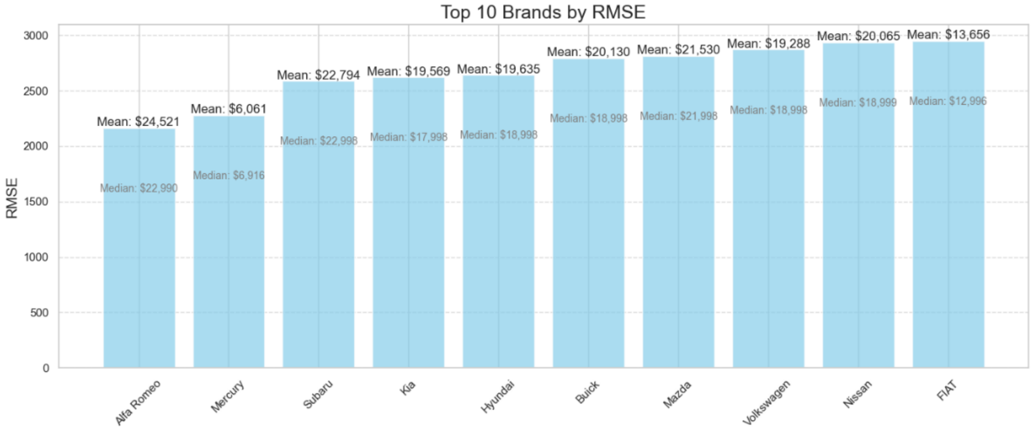
Visualization of Residuals



At 99% of Condifence Rate I checked whether residuals distributed by normal or not and the result is residuals are not distributed normally.

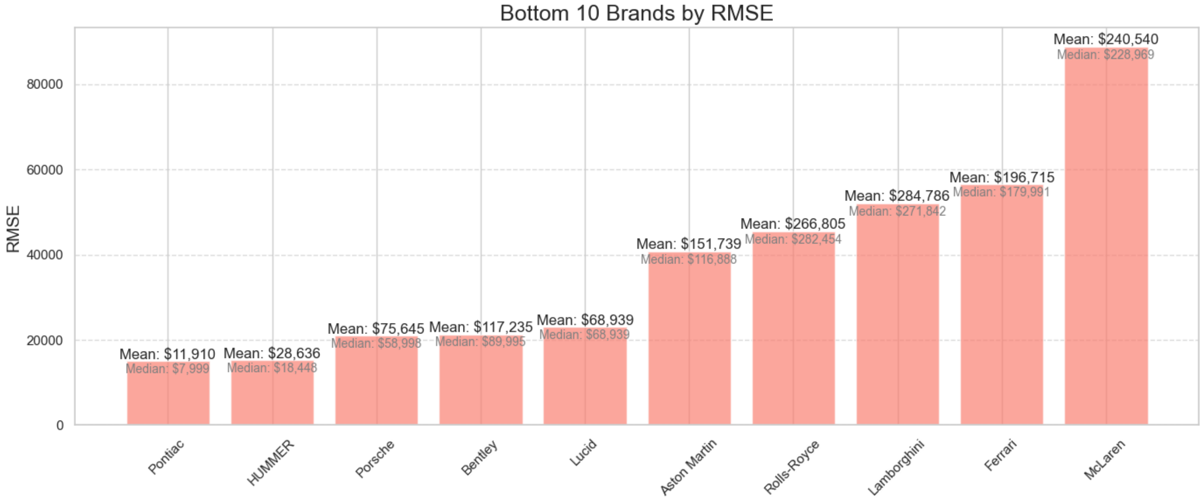
### Identifying Each Car Brands contribution to RMSE,MAE Scores

Car brands that have the lowest RMSE values



These 10 brands have the lowest RMSE values and we can certainly say that our model most likely to predict car prices of these car with small percentage of error.

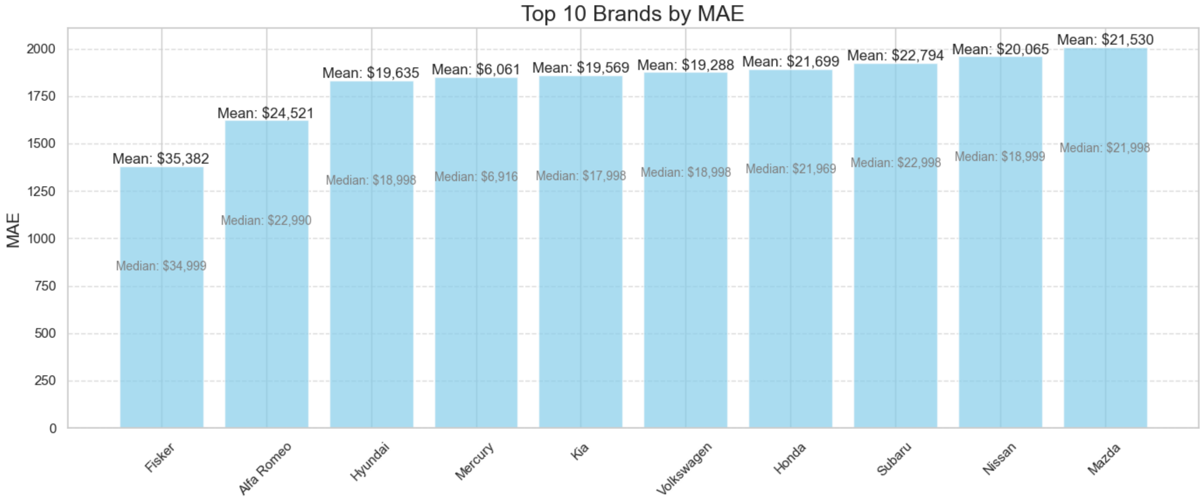
Car brands that have the Highest RMSE values



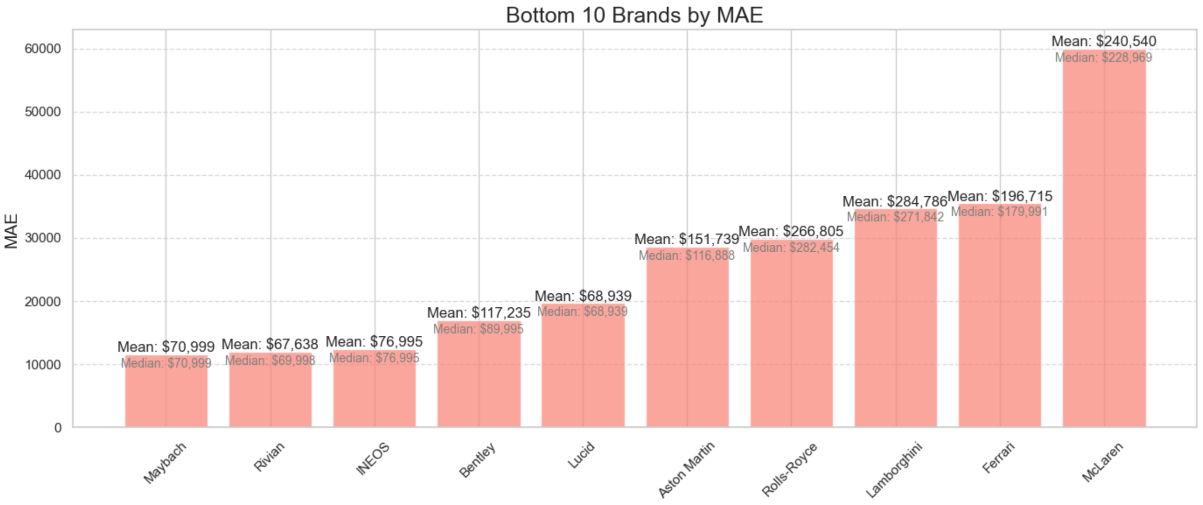
We can clearly see that more expensive car brands exhibit higher RMSE values due to their elevated prices and limited representation in the dataset. This is because supercars are often unique and produced in smaller quantities. Additionally, while Pontiac and HUMMER are relatively cheaper compared to other brands, they also have high RMSE scores. This is likely due to their limited presence in our dataset, which hinders the model's ability to fully understand the characteristics of these instances due to insufficient data.

Car brands that have the Highest RMSE values

Car brands that have the lowest MAE values.



Car brands that have the highest MAE values



As mentioned earlier, the same observation applies to the MAE metric. The combination of limited data and high prices results in greater model error for expensive car brands. Therefore, it's crucial to use these metrics with caution. Incorporating a confidence interval when predicting the prices of expensive cars could provide additional insights and enhance the reliability of our predictions.

## Conclusion

In this project, we achieved significant success in developing a machine learning model for predicting car prices, evidenced by a Root Mean Squared Error (RMSE) of 6,280 and a Mean Absolute Error (MAE) of 2,953. These metrics indicate that our model is capable of providing reasonably accurate price predictions, with the RMSE highlighting the average error magnitude and the MAE underscoring the model's robustness in capturing absolute deviations.

Moreover, our model attained an impressive $R^2$ score of 0.92, demonstrating that it can explain 92% of the variability in car prices. This level of performance signifies a strong correlation between the predicted and actual prices, which is crucial for stakeholders, such as dealerships and potential buyers, seeking to understand market dynamics and make informed purchasing decisions.

Throughout the project, we employed a variety of advanced techniques, including cross-validation and hyperparameter optimization with the Optuna library, to enhance model accuracy. Despite challenges, particularly with high-priced car brands, our findings reveal that the model effectively navigates the complexities of car pricing.

The high RMSE and MAE for expensive brands can be attributed to their limited representation in the dataset, which underscores the model's dependence on data quality and quantity. This observation highlights the need for careful consideration when interpreting model outputs, especially for unique vehicle categories like supercars.

In conclusion, the project not only demonstrates the power of machine learning in car price prediction but also sets the foundation for future enhancements. Strategies such as incorporating confidence intervals for expensive car predictions could further enrich the model's utility. Overall, our results affirm the viability of machine learning as a valuable tool in the automotive industry, providing insights that drive better decision-making for both consumers and sellers. Also keep in mind that prices may vary on the factors that are not considered in this dataset, the accidents of the car that suffered or fix counts of the parts etc.

This project has been developed by **Emir Tatlıcı.**

(Statistics Undergraduate 3rd Grade Student at Yıldız Technical University)

Project GitHub Link → https://github.com/Emirtatlici/End-to-End-ML-Regression-Price-Prediction

Linkedin → https://www.linkedin.com/in/emir-tatlici/