ESTIMATING PRICE REDUCTION FOR VEHICLES

Sabancı Üniversitesi FACULTY OF ENGINEERING AND NATURAL SCIENCES

Student(s)
Ege Alpay, Gizem Aydın, Ceren Çamlıbel, Ezgi Genç, Güniz Irmak Köksalan, Aynur Süne

Faculty Member(s)
Kamer Kaya, Öznur Taştan, Kemal Kılıç, Sinan Yıldırım

Company Advisor(s)
Senem Yıldırım

ABSTRACT

As fleet leasing is becoming more and more popular each day, determining the prices of the fleets is getting harder. The loss in value of vehicles during the leasing period plays a key role on determining contract prices. As this is an ever-growing area, the industry would like to automate this process.

The company that is being worked with, Doğuş Technology, has a customer fleet leasing company named VDF Filo. VDF Filo leases cars for an approximate period of 3 years and sells them in the second-hand market without adding a profit margin. That is why, knowing the value of vehicles after leasing period is very important for their company.

OBJECTIVES

The main objective of the project is to create an automated model based on machine learning methods to predict the prices of the second hand vehicles after their leasing period. To obtain more accurate predictions, macroeconomics factors were also considered.

PROJECT DETAILS

Steps of the project are described in the figure below.

Step 1: Cleaning & Understanding Data
- Data confirmed that the price of a car was directly proportional to its manufacturing year and reversely proportional to its kilometers.
- Sales data from 2014 to 2018 was cleaned individually, features as: horsepower, currency, damage, ÖTV, gearbox, fuel type, marka-model infos etc. were added. Then the datasets were merged together.

Step 2: Creating Baseline Models
- The average sale price of each model was extracted from the training set, and the corresponding average sales price was appended to the validation set as a prediction.
- The aim of this was to see how the simplest model predicts the price of a car.

Step 3: Outlier Elimination
- Outliers affect the machine learning models negatively. Outliers are some of the samples that have abnormal label values compared to other samples due to intrinsic errors.
- A few different ways (Interquartile Range Method, Z-score,Random Forest Regressor) were implemented on the training dataset to see whether outlier elimination had any impact on the performance of the models. After these results were obtained, it was decided that no outlier elimination technique would be used as there was no performance gain.

Step 4: Model Testing
Models using several machine learning algorithms such as XGBoost, Random Forest, Neural Network and Lasso Regularized Regression were generated. The results were compared using error metrics.

MODELS & METHODS

During this project, four machine learning algorithms were used to create models.

Random Forest	Forms an ensemble of decision trees and trains them using bagging.
Neural Network	Adapts the structure of interconnected group of artificial neurons to minimize prediction error. Dropout layers were used for regularization and feature embedding layers were used for categorical fields.
Lasso Regression	Regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the model.
XGBoost	Implementation of gradient boosted decision trees designed for speed and performance.

Macroeconomic factors: Increase in inflation results with a decrease in purchasing power which affects second-hand automotive industry as well. Using currency and ÖTV instead of inflation would be more beneficial for the project. Since past data for monthly currency and ÖTV changes were easier to observe and implement into model, a column for each feature is added to the project.

Main (Error) Metrics:

Mean Absolute Percentage Error (MAPE): For each sample, the absolute error percentage is calculated; and these values are averaged over each sample.

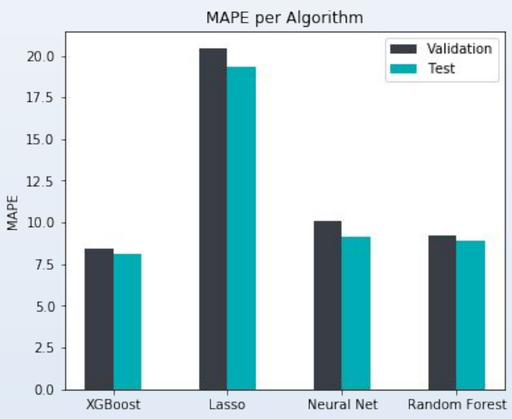
Mean Absolute Error (MAE): Average of the absolute value of the difference between the true value and the predicted value.

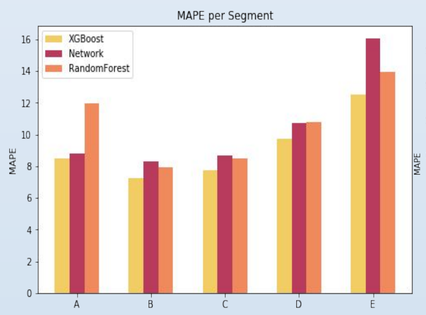
R Squared: This value shows how much of the variance can the model catch.

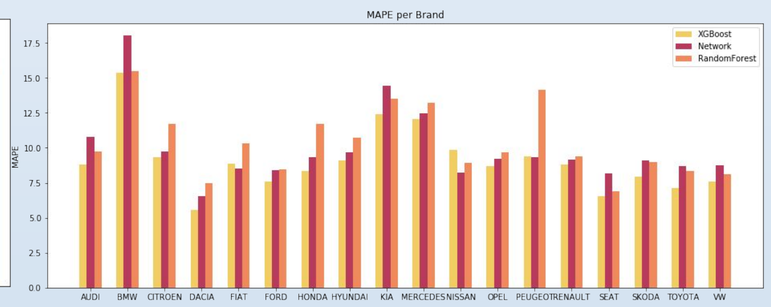
RESULTS

PROBLEM 1

- The datasets having the sales data from the year 2014 to 2018 was merged, and then divided into training, validation and testing datasets having 60%, 20% and 20% of the samples respectively.
- Splitting was done in a stratified fashion with respect to months.
- This is the easiest problem, and useful for the company DOD, who sells second hand cars.
- Best result was obtained by XGBoost, MAE is ₺5,253 on test set.

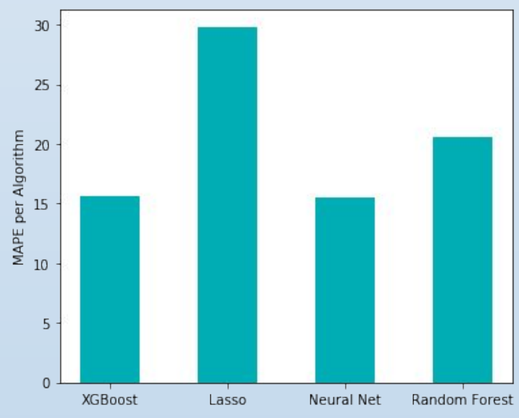


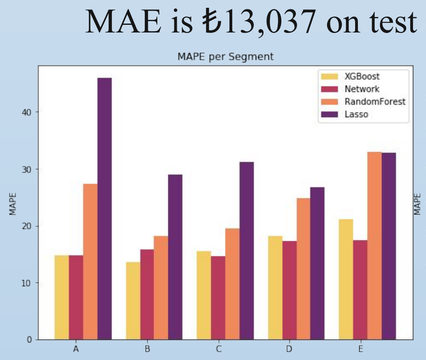


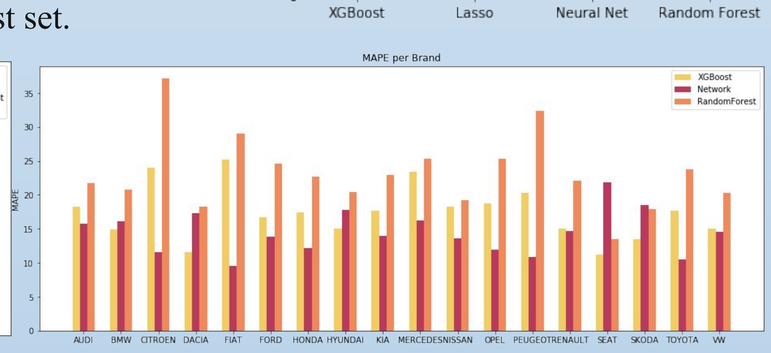


PROBLEM 2

- Training Set includes sales from years 2014 and 2015.
- Test Set includes sales from years 2017 and 2018.
- The aim is to see which algorithm can predict the future prices.
- This problem is harder than the first one, and useful for VDF.
- Best result was obtained by XGBoost, MAE is ₺13,037 on test set.

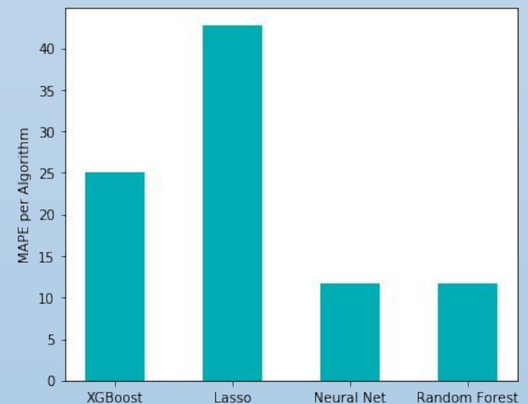


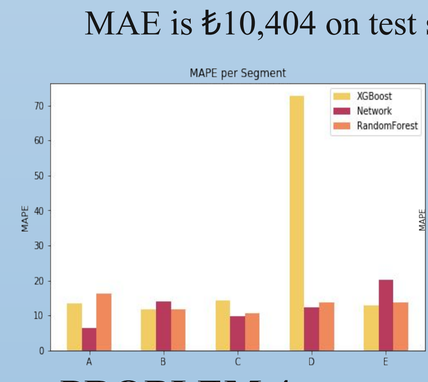


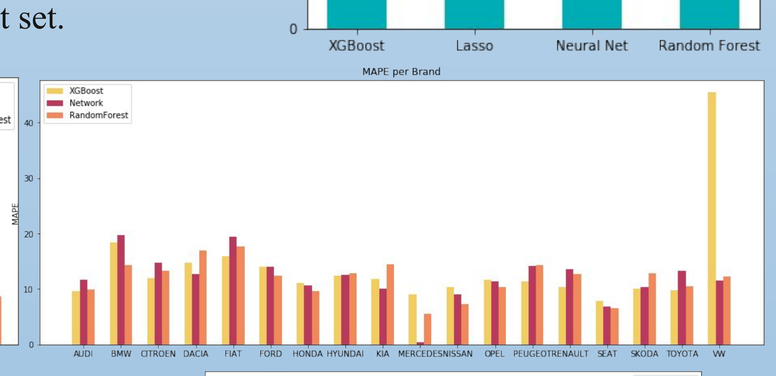


PROBLEM 3

- Training Set includes all data from 2014 to 2018, approx. 80k samples.
- Test Set includes sales data of 2019, approx. 2000 samples.
- This is also a hard problem as it requires predicting the future prices.
- Best result was obtained by XGBoost, MAE is ₺10,404 on test set.

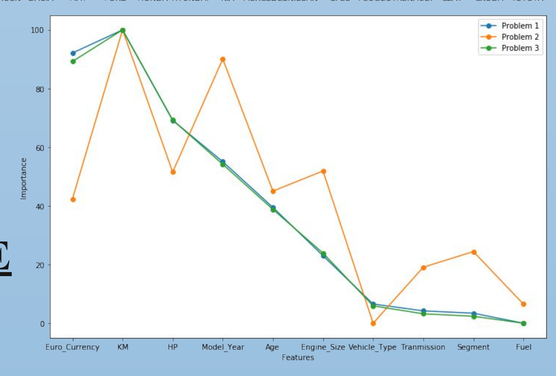






PROBLEM 4

In this part of the project, aim was to compare Feature Importance of XGBoost between first 3 problems.



CONCLUSION & FUTURE WORK

A Graphical User Interface will be implemented so that VDF can use XGBoost algorithm to predict the second hand price of the fleets they lease.

REFERENCES

Srinivasan A., Kamble G., Bhartiye A. (2007). Estimating Vehicle Prices Using Market Data. USA

Guo, C., Berkhahn F. (2016). Entity Embeddings of Categorical Variables. arXiv:1604.06737 [cs.LG]