

**Submitted By**

Ahmad Awais (10931)

Muhammad Osama (10950)

BS-SE (7th)

**Submitted To**

Dr Bilal Khan

**Department of Computer Science**

**City University of Science and Information Technology**

**Peshawar – Pakistan**

**Introduction**

We Determining the listed price of a used car. Our goal is to predict the cars price accurate. When we buy a car, we check the prices of different cars online and sellers put different prices so we cannot fully trust on sellers either they put the right price or not. We develop four models using old data of the cars. Our focus of this project is developing machine learning models that can accurately predict the price of a used car based on its features, in order to make informed purchases. We implement and evaluate various learning methods on a dataset. In our data set there are 4345 Rows and 9 columns.

Columns include:

1. Brand
2. Price
3. Body
4. Mileage
5. EngineV
6. Engine Type
7. Registration
8. Year
9. Model

The algorithms we used for the projects are Random Forest, Decision Tree, Gradient Boost and XGBRegressor the reason of us

**Problem Statement**

Deciding whether a used car is worth the posted price when you we see online cars on OLX and different Apps it is difficult to guess the accurate price of the car. Several factors, including mileage, make, model, year, etc. can influence the actual worth of a car. From the perspective of a seller, it is also a dilemma to price a used car appropriately. Based on existing data, our goal is to use machine learning algorithms and preprocessing steps to develop models for predicting used car prices.

**Experimental Setup**

We utilized several classes and state-of-the-art methods, including ensemble learning techniques, with 80% - 20% split for the training and test data. Linear Regression, Random Forest and Gradient Boost were our baseline methods. For most of the model implementations the open-source Scikit-Learn package was used.

In preprocessing steps

* We handle missing values
* Remove outliers
* Convert categorical data to Onhotencoding

**Libraries**

Pendas

NumPy

Matplotlib

Seaborn

Sklearn

**Tool**

Jupyter Notebook

**Results Analysis and Discussion**

In this project we make four Machine Learning Models and Our results show that Random Forest, XGBRegressor, Decision Tree and Gradient Boost yield the best results, but are compute heavy. By comparative analysis we find that XGBRegressor has the less mean absolute error because xgboost is powerful algorithm for continuous data

We got best result in all models one of the reasons is best preprocessing steps. We took basics and important preprocessing steps for our model

**Conclusion**

Mean Absolute Error of Random Forest = 2687.4582921614488

Mean Absolute Error of Decision Tree = 2980.4198100335852

Mean Absolute Error of Gradient Boost = 3075.4882499940823

Mean Absolute Error of Xgboost = 2404.7055743946294

**References**

1. https://www.kaggle.com/jpayne/852k-used-car-listings

2. N. Monburinon, P. Chertchom, T. Kaewkiriya, S. Rungpheung, S. Buya and P. Boonpou, "Prediction of prices for used car by using regression models," 2018 5th International Conference on Business and Industrial Research (ICBIR), Bangkok, 2018, pp. 115-119.

3. Listiani M. 2009. Support Vector Regression Analysis for Price Prediction in a Car Leasing Application. Master Thesis. Hamburg University of Technology

4. Ke, Guolin, et al. "Lightgbm: A highly efficient gradient boosting decision tree." Advances in Neural Information Processing Systems. 2017.

5. Fisher, Walter D. "On grouping for maximum homogeneity." Journal of the American statistical Association 53.284 (1958): 789-798.

6. https://scikit-learn.org/stable/modules/classes.html: Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.