Used Car Price Prediction

Abstract:

The goal of this project was to use regression models to predict the price of used car in order to help improve fair sell. The used data in this project is provided by Kaggle. With sklearn library random forest was trained and get 88% accuracy. After refining a model, I built an interactive visualization and communicate my results using seaborn library.

Design

This project is one of the T5 Data Science BootCamp requirements. The main aim of this project is to predict the price of used car. Data provided by Kaggle has been used in this project. All the Lifecycle In A Data Science Project is divided into four parts:Exploratory Data Analysis, Feature, Engineering, Feature selection, Model.

Data

The dataset is provided in .csv format. It contains over 6000 rows with 13 featuers that contains information about used cars. The columns in the given dataset are as follows: (Name, Location, Year, Kilometers_Driven, Fuel_Type, Transmission, Owner_type, Mileage, Power, Engine, New_Price, Price, Seats).

Algorithms

- **Feature Engineering:**

1. Adding a new column in our data frame : The ‘Name’ column has so many values so we will separate the brand names from the column and create a new column ‘Brand_Name’.

2. Handling of Missing values : Removing null values from the 3 columns , fill null value in column seat by mode.

3. Substitution of Categorical variables : I used one hot encoding. This essentially turns every unique value of a variable into its own binary variable.

4. Remove Outliers

- **Model**

Linear Regression,

Lasso

, Decision Tree,

Random Forest

were used before settling on random forest as the model with best accuracy. Random forest feature importance ranking was used directly to guide the choice and order of variables to be included as the model underwent refinement.

Model Evaluation and Selection

Random Forest get the higher accuracy get 88%. The models were trained on a 20/80 test vs

---

Tools

1. Pandas for data manipulation
2. Scikit-learn for modeling
3. Matplotlib for plotting
4. Numpy.
5. The work will be done through Jupyter Notebook.