Bijay Adhikari

Learning Journals and Findings in Project-3:

Title:

Used Car Price Prediction

Dataset:

The dataset was taken from Open Data Nepal. The dataset consists of 2217 rows and 12 columns. There are 11 independent features and 1 label feature i.e Price.

Description of Dataset:

1. Brand: The brand of the car
2. Model: Car's model
3. Model_year: The year when the car was manufactured
4. Transmission: The transmission allows the vehicle to change gears,
5. Engine_size(cc): the size of an engine refers to the total volume of air and fuel being pushed through the engine by the cylinders. It's measured in cubic centimetres (cc).
6. Drivetrain: the system in a motor vehicle which connects the transmission to the drive axles. 2WD-Two Wheel Drive and 4WD-Four Wheel Drive
7. Fuel_type: Car fuel type-Petrol or diesel
8. Colour: Color of the car
9. Lot_no: A lot number is an identification number assigned to a particular quantity like a car.
10. Kilometer: distance covered in kilometer
11. Status: used or new car
12. Price: price of the car

Scope:

There is a separate market for the buy and sell of used cars. Lots of people buy and sell cars for various reasons. It is difficult for dealers to put a price on the cars without looking for the manufacturer company, car model, and the other characteristics.

Problem Statement: To build a model that can predict the price of a new coming used car model.

I have cover following topics in my project:

1. Statistical Analysis
2. Handling of missing values
3. Handling of outliers
4. Data visualization and findings
5. Feature engineering
6. Multiple Model Building and Evaluation

Limitation:

1. Dataset has more number of one class and less number of other classes in some features. Data is imbalanced not in target variables but other independent categorical variables.
2. The available data has 12 features including Price. Some of the features like size of car (no of seat), condition of car (very good, good, bad, worst) etc are missing. There are many factors to influence in pricing of cars so if all are not included then it could not be an accurate model in real scenario.
3. The dataset does not cover all classes even in those included features. Like for transmission feature , there are only 2 classes:
    - Automatic Transmission (AT)
    - Manual Transmission (MT)

    But in reality we have car with 4 classes/types of transmission:
    - Automatic Transmission (AT)
    - Manual Transmission (MT)
    - Automated Manual Transmission (AM)
    - Continuously Variable Transmission (CVT)

Hypothesis:

1. The machine learning model is built based on available dataset. So the model is learning only those features that are mentioned in the dataset. It fails if any feature not in the dataset is tested on it.
2. Cars have many brands and models (>100)so it is difficult to include all classes in one hot encoding so only top 15 models are taken into consideration and considered all as a zero.

Findings:

1. In the dataset, there are equal no of missing values in all columns of the dataset. But when I saw the dataset I found that these are not missing values but empty rows. That's why all missing values are equal so we remove these empty rows.
2. If mean and median of data is same that doesn't always mean the dataset is symmetrical. Sometimes the outliers shift the mean closer to median but when those outliers are removed then the distribution is either left or right skewed.
3. There is an empirical formula in Gaussian distribution (mean-3sd, mean, mean+3sd). If data does not fall in this range then it is consider outlier in gaussian distribution
4. Data is not just about analysis and visualization. But mostly we have to relate to real world scenarios. Like in my data analysis the Kilometer travelled and price were negatively correlated. This makes sense because the car which has travelled more distance is supposedly old and has a lower price.
5. To handle multiple classes of categorical data, I have used top 15 dummy variables one hot encoding but there are other techniques like Label Encoder.

6. Features : Model year and Lot no are negatively strongly correlated as both represent the age of the vehicle but in opposite ways. Model year gives the age of the car and Lot no also defines the age of the car.
7. The maximum time I was stuck while doing this project was whether to keep color as features or not. It does seem to be true that color can really impact the price of the car. If I am going to buy a car I would only prefer matte black or grey silver. And that is me as a buyer and my preference. I found out that color can impact the sales rate. Like cars with preferred color are sold faster than others but does it increase the price of the car? Generally no.
8. Decision, RandomForest and GradientBoosting algorithms model works better than linear regression or any other model. Because of following reasons:

    - Ensemble model is not affected by outliers present in the dataset
    - Ensemble model even performs great in unbalanced dataset
    - Ensemble model is not affected by missing values.