CARS PRICE PREDICTION PROJECT

Submitted by:

ABHILASHA MEWADA

INTRODUCTION

- Business Problem Framing

The production of cars has been steadily increasing in the past decade, with over 70 million passenger cars being produced in the year 2016. This has given rise to the used car market, which on its own has become a booming industry. The recent advent of online portals has facilitated the need for both the customer and the seller to be better informed about the trends and patterns that determine the value of a used car in the market. Using Machine Learning Algorithms such as Lasso Regression, Multiple Regression and Regression trees, we will try to develop a statistical model which will be able to predict the price of a used car, based on previous consumer data and a given set of features.

We are required to model the price of cars with the available independent variables. It will be used by the management to understand how exactly the prices vary with the independent variables. They can accordingly manipulate the design of the cars, the business strategy etc. to meet certain price levels. Further, the model will be a good way for management to understand the pricing dynamics of a new market..

- Review of Literature:

The automotive industry is composed of a few top global

Multinational players and several retailers. The multinational

Players are mainly manufacturers by trade whereas the retail market features players who deal in both new and used Vehicles. The used car market has demonstrated a significant Growth in value contributing to the larger share of the overall Market. The used car market in India accounts for nearly 3.4 Million vehicles per year

Cars24 is a web platform where seller can sell their used car. It is an Indian Start-up with a simplified user interface which Asks seller parameters like car model, kilometres travelled, Year of registration and vehicle type (petrol, diesel)[1]. These allow the web model to run certain algorithms on given Parameters and predict the price.

Analytical Problem Framing

- Mathematical/ Analytical Modelling of the Problem

Python was the major technology used for the Implementation of machine learning concepts the reason being that there are numerous inbuilt methods in the form of packaged libraries present in python. Following are Prominent libraries/tools we used in our project.

Numpy is a general-purpose array-processing package. it provides a high-performance multidimensional array object and tools for working with these arrays. It is the fundamental package for scientific computing with Python. Besides its obvious scientific uses, NumPy can also be used as an efficient multi-dimensional container of generic data.

Arbitrary data-types can be defined using Numpy which allows NumPy to seamlessly and speedily integrate with a wide variety of databases.

SciPy is a free and open-source Python library used for scientific computing and technical computing. SciPy contains modules for optimization, linear algebra, integration, interpolation, special functions, FFT, signal and image processing, ODE solvers and other tasks common in science and engineering.

SciPy builds on the NumPy array object and is part of the NumPy stack which includes tools like Matplotlib, pandas.

- Data Sources and their formats

Data is scraped form the cars24 and all data is of used cars in india from different cities. Than it stored in the form of csv and we can also later use it. Data frame has 11 columns and 5521 rows.

- Data Pre-processing Done

Data pre-processing done with many steps, first we check info than we check the null values. it has many null values than we just remove them because it can affect the performance of the model.

We have describe the data than checked the correlation between the columns. We have checked outliers also but we have not removed it because it can affect the prediction.

We have transformed the data with standard scaler transform and split the variables.

Encoding:

Label Encoding: is a popular encoding technique for handling categorical variables. In this technique, each label is assigned a unique integer based on alphabetical ordering.

Label Encoding challenges: there is a very high probability that the model captures the relationship between values like they were ordinal which isn't suitble for example for ocean proximity here.

Outliers:

"An outlier is an observation that lies an abnormal distance from other values in a random sample from a population. In a sense, this definition leaves it up to the analyst (or a consensus process) to decide what will be considered abnormal. Before abnormal observations can be singled out, it is necessary to characterize normal observations. Two activities are essential for characterizing a set of data of the overall shape of the graphed data for important features, including symmetry and departures from assumptions of the data for unusual observations that are far removed from the mass of data. These points are often referred to as outliers. Two graphical techniques for identifying outliers, scatter plots and box

plots, along with an analytic procedure for detecting outliers when the distribution is normal (Grubbs' Test)"}]

Summary of Data Pre-processing:

The data went through the process of cleaning as follows:

Explored and visualized the data Handled categorical data encoding using label encoder which is more suitable for the case

- Hardware and Software Requirements and Tools Used

importing essential libraries

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import StandardScaler,MinMaxScaler,power_transform
from sklearn.linear_model import LinearRegression,LogisticRegression
from sklearn.metrics import r2_score, mean_squared_error,accuracy_score,roc_auc_score,roc_curve,confusion_matrix,classification_report
from sklearn.ensemble import AdaBoostRegressor,
```

```
import warnings
warnings.filterwarnings('ignore')
from sklearn import preprocessing
import scipy.stats as stats
from scipy.stats import zscore
from sklearn.svm import SVC
from sklearn.naive_bayes import MultinomialNB
```

Model/s Development and Evaluation

The process of modeling means training a machine learning algorithm to predict the labels from the features, tuning it for the business need, and validating it on holdout data. The output from modeling is a trained model that can be used for inference, making predictions on new data points.

A machine learning model itself is a file that has been trained to recognize certain types of patterns. You train a model over a set of data, providing it an algorithm that it can use to reason over and learn from those data. Once you have trained the model, you can use it to reason over data that it hasn't seen before, and make predictions about those data. For example, let's say you want to build an application that can recognize a user's emotions based on

their facial expressions. You can train a model by providing it with images of faces that are each tagged with a certain emotion, and then you can use that model in an application that can recognize any user's emotion

Linear regression:

Linear regression is a supervised machine learning method that is used by the Train Using AutoML tool and finds a linear equation that best describes the correlation of the explanatory variables with the dependent variable.

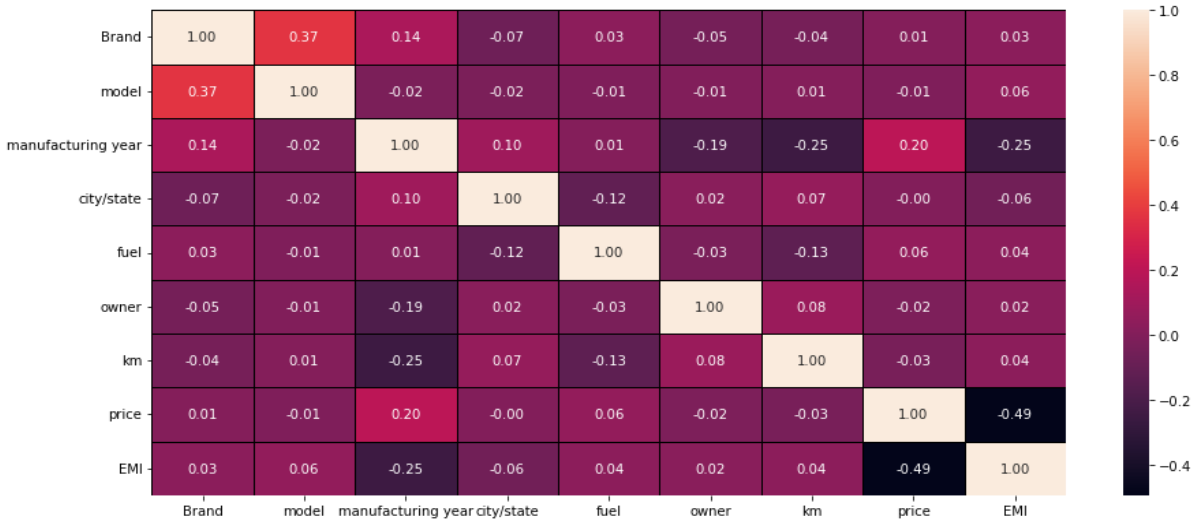
Lasso regression:

The Lasso class takes in a parameter called alpha which represents the strength of the regularization term. A higher alpha value results in a stronger penalty, and therefore fewer features being used in the model. In other words, a higher alpha value such as 1.0 results in more features being removed from the model than a value such as 0.1. The Lasso class also has a fit() method that can be used to fit the model to training data, and a predict() method that can be used to make predictions on new data.

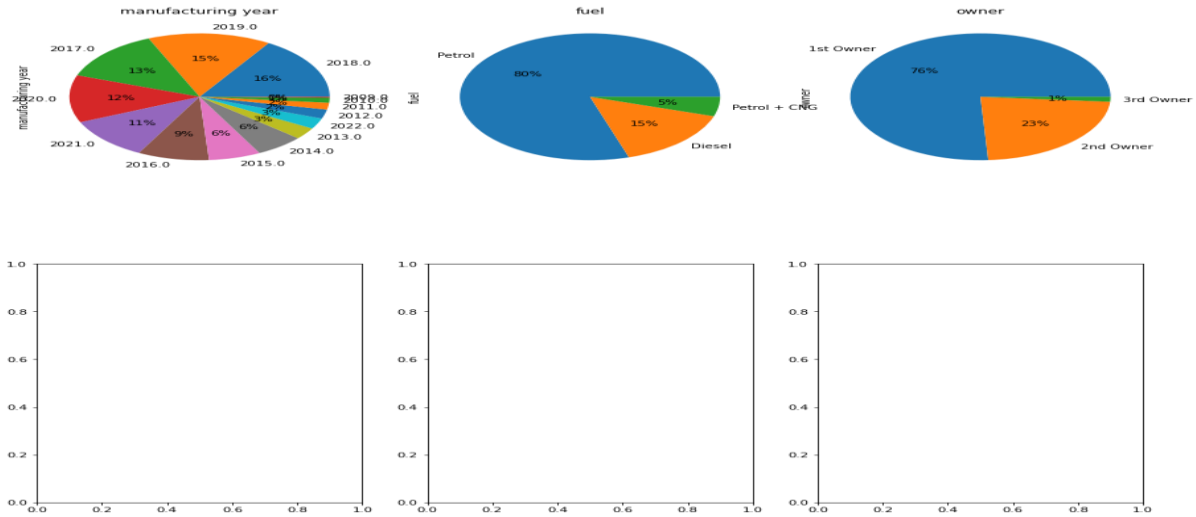
Ridge regression:

Ridge regression is used to solving the problem of multicollinearity when the independent variables are highly correlated with each other, and the correlation matrix will be singular and we can’t obtain a unique parameter.

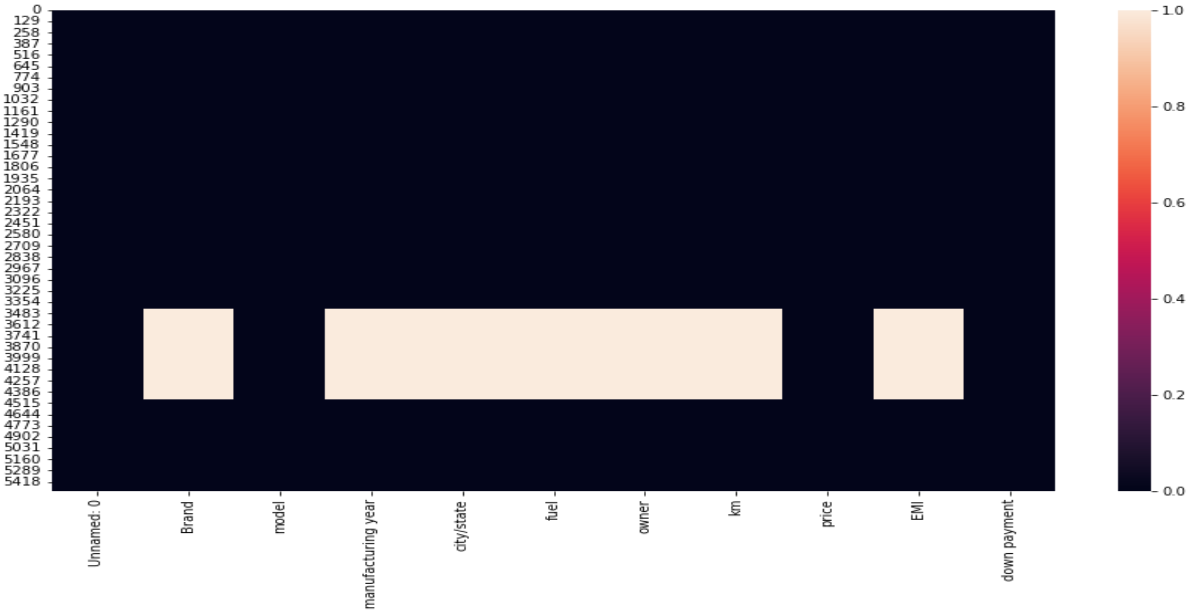
Visualisation:



This shows that the data columns are not all correlated to each other.



It shows that the cars which are manufactured in 2018 and 2019 are highly selling cars. Most of the cars have petrol fuel. The 1st owner cars are mostly are on sale.



Dataframe has some null values which we have to handle.

Conclusions:

1. We used several regression models to fit our data and it seems that they all succeeded to fit the data well and this indicated that the data pre-processing stage was also a success but we're still facing the problem of overfitting so I see that all models are truly promising and ready for the next stage of improvement to reduce overfitting.

2. Adaboost regression give the best regression model for cars price prediction.