**Project Report**

**On**

# Used Car Price Prediction



Submitted in partial fulfilment for the award of

Post Graduate Diploma in Big Data Analytics (PG-DBDA)

From Know- IT (Pune)

**Guided by:**

**Mr. Amey Manjrekar**

## Submitted By:

Hirkani Badgujar (210943025007)

Gayatri Pendkar (210943025015)

Rahul Gitte (210943025016)

Shubham Katkar (210943025022)

**CERTIFICATE**

**TO WHOMSOEVER IT MAY CONCERN**

**This is to certify that**

Hirkani Badgujar (210943025007)

Gayatri Pendkar (210943025015)

Rahul Gitte (210943025016)

Shubham Katkar (210943025022)

**Have successfully completed their project on**

# Used Car Price Prediction

# 

**Under the guidance of Mr. Amey Manjrekar**

**ACKNOWLEDGEMENT**

This project **“Used Car Price Prediction”** was a great learning experience for us and we are submitting this work to CDAC Know-IT (Pune).

We all are very glad to mention the name of **Mr. Amey Manjrekar** and **Mr. Sanjay Sane** for his valuable guidance to work on this project. His guidance and support helped us to overcome various obstacles and intricacies during the course of project work.

We are highly grateful to **Mr. Vaibhav Inamdar** Manager (Know-IT), C-DAC, for his guidance and support whenever necessary while doing this course Post Graduate Diploma in Big Data

Analytics (PG-DBDA) through C-DAC ACTS, Pune.

Our most heartfelt thanks goes to **Mrs. Bakul Joshi** (Course Coordinator, PG-DBDA) who gave all the required support and kind coordination to provide all the necessities like required hardware, internet facility and extra Lab hours to complete the project and throughout the course up to the last day here in C-DAC Know-IT, Pune.

**From:**

Hirkani Badgujar (210943025007)

Gayatri Pendkar (210943025015)

Rahul Gitte (210943025016)

Shubham Katkar (210943025022)

**TABLE OF CONTENTS**

**ABSTRACT**

**1. INTRODUCTION**

#### 2. SYSTEM REQUIREMENTS

**2**.1 Software Requirements

2.2 Hardware Requirements

**3. FUNCTIONAL REQUIREMENTS**

#### 4. SYSTEM ARCHITECTURE

1. **METHODOLOGY**

#### 6. DATA VISUALIZATION AND REPRESENTATION

#### 7. RESULT AND FINDING

#### 8. CONCLUSION AND FUTURE SCOPE

**References**

## ABSTRACT

Determining whether the listed price of a used car is a challenging task, due to the many factors that drive a used vehicle’s price on the market. The 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 consisting of the sale prices of different makes and models across cities in the United States.

Initially we downloaded raw data from <https://github.com/abhashpanwar/used-car-price-prediction> We are processing ETL operation on top of that data using Apache Spark and dumping required data into Mongo Db for data visualization through PowerBI and build Machine Learning to predict price of used cars. Precise estimation of price supports the decision-making process means it helps the seller and customer to decide the price of used car who use such systems.

1. **INTRODUCTION**

The price of a new car in the industry is fixed by the manufacturer with some additional costs incurred by the Government in the form of taxes. So, customers buying a new car can be assured of the money they invest to be worthy. But, due to the increased prices of new cars and the financial incapability of the customers to buy them, Used Car sales are on a global increase. Therefore, there is an urgent need for a Used Car Price Prediction system which effectively determines the worthiness of the car using a variety of features.

Existing System includes a process where a seller decides a price randomly and buyer has no idea about the car and its value in the present day scenario. In fact, seller also has no idea about the car’s existing value or the price he should be selling the car at. To overcome this problem, we have developed a model which will be highly effective. Regression Algorithms are used because they provide us with continuous value as an output and not a categorized value. Because of which it will be possible to predict the actual price a car rather than the price range of a car

In order to predict price of used car, data which would be available before the sale or purchase of the car. This includes manufacturer, condition, engine capacity, odometer, latitude, longitude etc.

#### Datasets and features:

Data used in the project is unstructured in nature from year 1995. It was collected from <https://github.com/abhashpanwar/used-car-price-prediction>. Linear Regression with model selection, Lasso, Ridge, AdaBoost, XGBoost, Bagging Regressor and Random Forest models were used to predict price for used car.

**2. SYSTEM REQUIREMENTS**

#### Hardware Requirements:

* Platform – Windows 10
* RAM – 8 GB of RAM,
* Peripheral Devices – Mouse, Keyboard, Monitor
* A network connection for data recovering over network.

#### Software Requirements:

* Python 3
* Apache Spark
* MongoDB
* PowerBI

####  OS – Window

**3. FUNCTIONAL REQUIREMENTS**

#### Python 3:

* Python is a general purpose and high level programming language.
* It is use for developing desktop GUI applications, websites and web applications.
* Python allows to focus on core functionality of the application by taking care of common programming tasks.
* Python is derived from many other languages, including ABC, Modula-3, C, C++, Algol-68,
* Small Talk, and UNIX shell and other scripting languages.

#### Apache Spark:

#### 

* Apache Spark is an open-source cluster computing system that provides high-level API in Java, Scala, Python and R.
* Apache Spark is one of the fastest-growing big data projects in the history of the Apache Software Foundation. With its memory-oriented architecture, flexible processing libraries, and ease-of-use, Spark has emerged as a leading distributed computing framework for real-time analytics.
* Spark is used for many types of data processing – it comes packaged with support for machine learning, interactive queries (SQL), statistical queries with R, graph processing, ETL, and streaming.
* For loading and storing data, Spark integrates with a number of storage MongoDB, and more.

#### MongoDB:

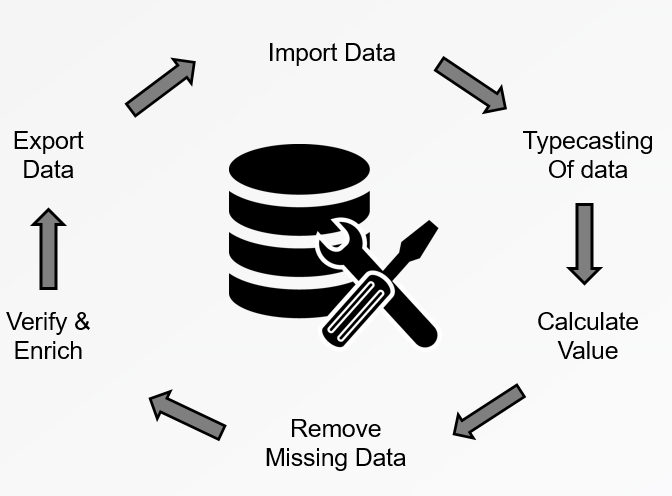
#### 

* **MongoDB**, the most popular NoSQL database, is an open-source document-oriented database.
* MongoDB allows a highly flexible and scalable document structure.
* MongoDB has built in solution for partitioning and sharing your database.
* MongoDB provides a variety of storage engines, allowing you to choose one most suited to your application.
* A real-life scenario for this kind of data manipulation is storing and querying real-time, intraday market data in MongoDB.

#### PowerBI:

* Data visualization is the graphical representation of information and data.
* It helps create interactive elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.
* PowerBI is widely used for Business Intelligence but is not limited to it.
* It helps create interactive graphs and charts in the form of dashboards and worksheets to gain business insights.
* All of this is made possible with gestures as simple as drag and drop.

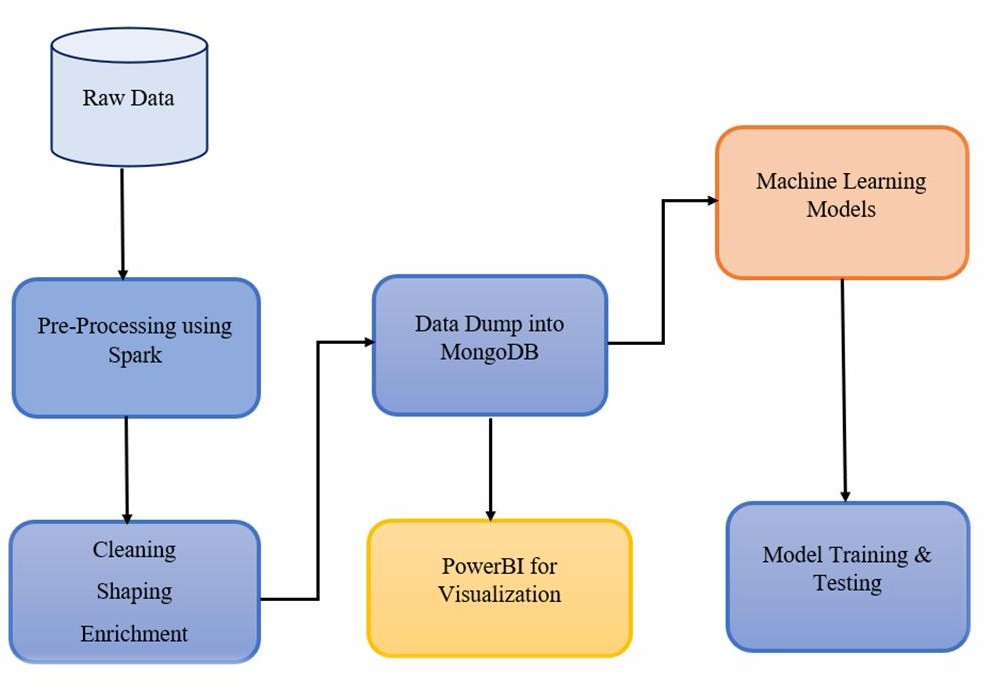
#### Data Cleaning Process:



**Fig: Data Cleaning Process**

Data cleansing or data cleaning is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data. Data cleansing may be performed interactively with data wrangling tools, or as batch processing through scripting. After cleansing, a data set should be consistent with other similar data sets in the system. The inconsistencies detected or removed may have been originally caused by user entry errors, by corruption in transmission or storage, or by different data dictionary definitions of similar entities in different stores. Data cleaning differs from data validation in that validation almost invariably means data is rejected from the system at entry and is performed at the time of entry, rather than on batches of data.

### 4. SYSTEM ARCHITECTURE

****

**Fig: System Architecture of Used Car Price Prediction.**

**5. METHODOLOGY**

In this project we have applied various different types of Regression

* Linear Regression
* Lasso Regression
* Ridge Regression
* KNN Regression
* Random Forest Regression
* Bagging Regression
* AdaBoost Regression
* XGBoost Regression

During the implementation we analyze the accuracy of all the algorithms.

**FEATURE SELECTION**

**Linear regression**

In linear regression, the relationships are modeled using linear predictor functions whose unknown model parameters are estimated from the data. The representation is a linear equation that combines a specific set of input values (x) the solution to which is the predicted output for that set of input values (y). As such, both the input values (x) and the output value are numeric.

**Lasso Regression**

The Lasso Regression is type of regularization (L1) can lead to zero coefficients i.e. some of the features are completely neglected for the evaluation of output.

Lasso regression not only helps in reducing over-fitting but it can help us in feature selection.

**Ridge Regression**

Ridge regression is a model tuning method that is used to analyse any data that suffers from multicollinearity. This method performs L2 regularization. When the issue of multicollinearity occurs, least-squares are unbiased, and variances are large, this results in predicted values being far away from the actual values.

**KNN Regression**

KNN regression is a non-parametric method that, in an intuitive manner, approximates the association between independent variables and the continuous outcome by averaging the observations in the same neighborhood.

**Random Forest**

Random forests are one the most popular machine learning algorithms.

Random forests consist of 4 –12 hundred decision trees, each of them built over a random extraction of the observations from the dataset and a random extraction of the features.

**Bagging Regression**

Bootstrap aggregating, also called bagging (from bootstrap aggregating), is a machine learning ensemble meta-algorithm designed to improve the stability and accuracy of machine learning algorithms used in statistical classification and regression.

It also reduces variance and helps to avoid overfitting.

**AdaBoost**

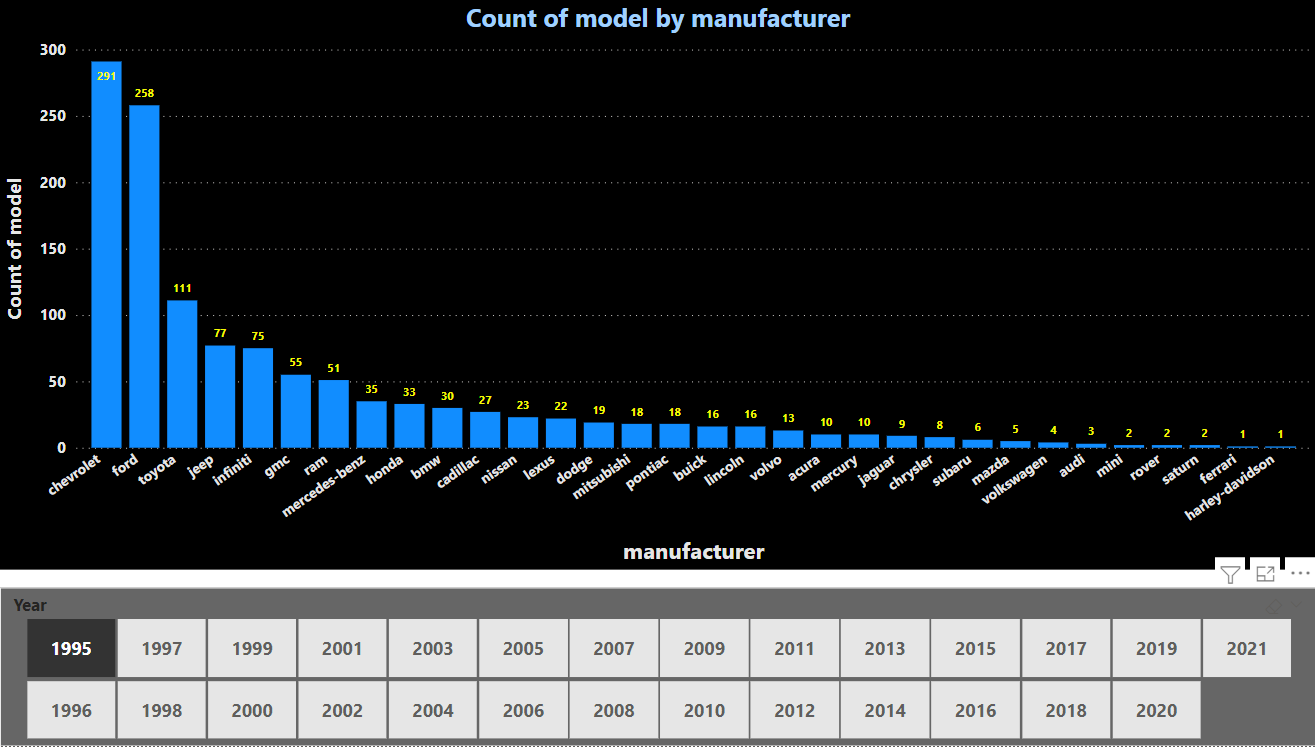
An AdaBoost regressor is a meta-estimator that begins by fitting a regressor on the original dataset and then fits additional copies of the regressor on the same dataset but where the weights of instances are adjusted according to the error of the current prediction.

**XG Boosting**

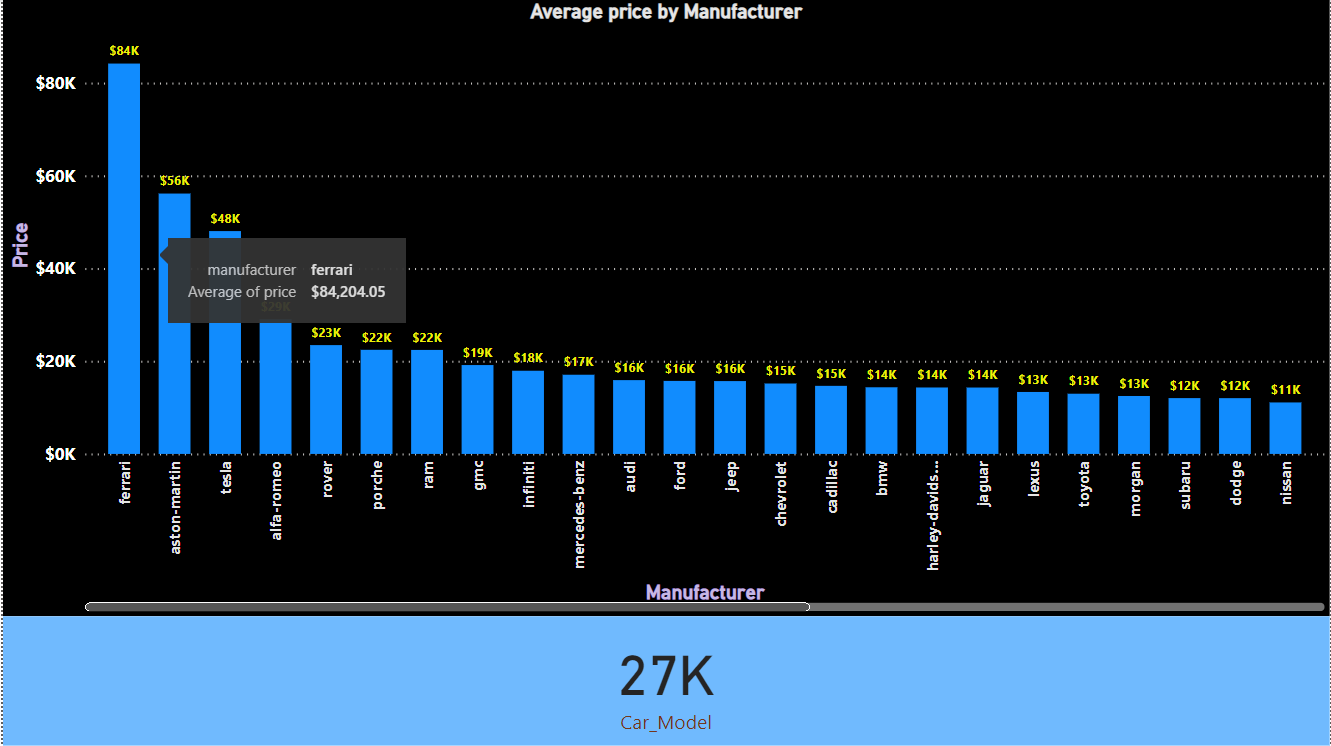
XG Boost is a powerful approach for building supervised regression models.

The validity of this statement can be inferred by knowing about its (XG Boost) objective function and base learners.

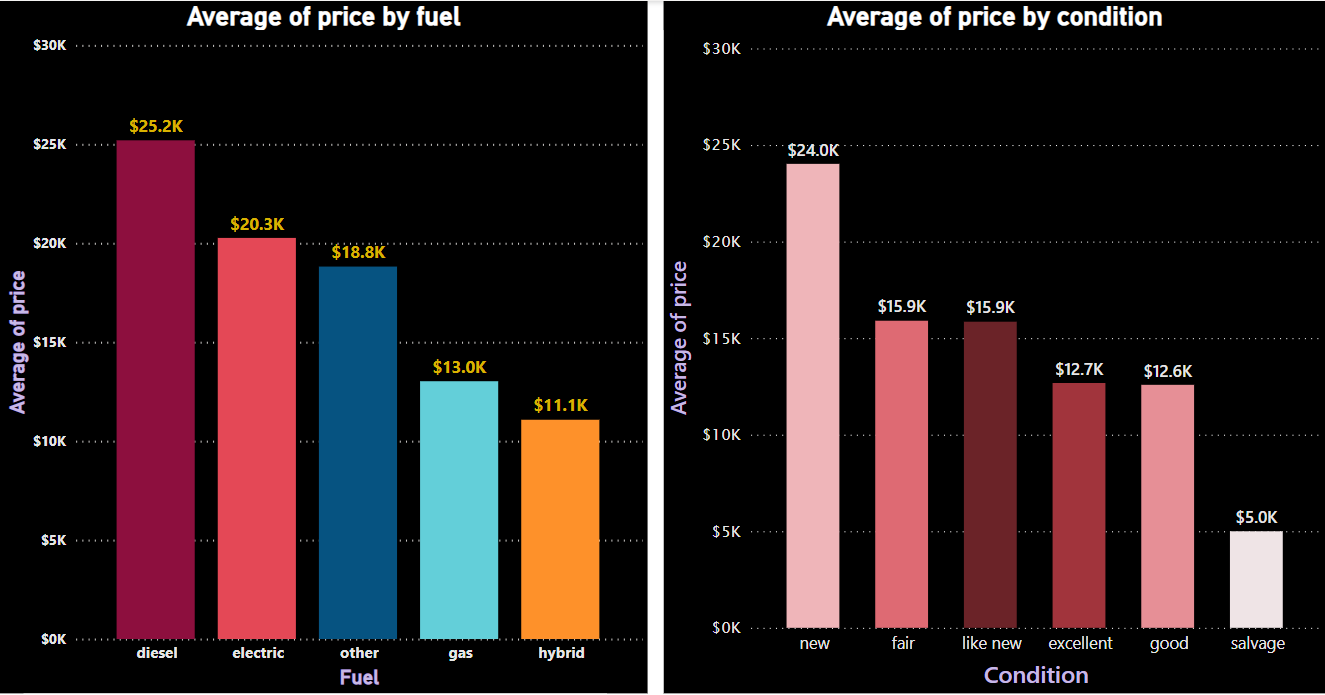
### 6. DATA VISUALIZATION AND REPRESENTATION

****

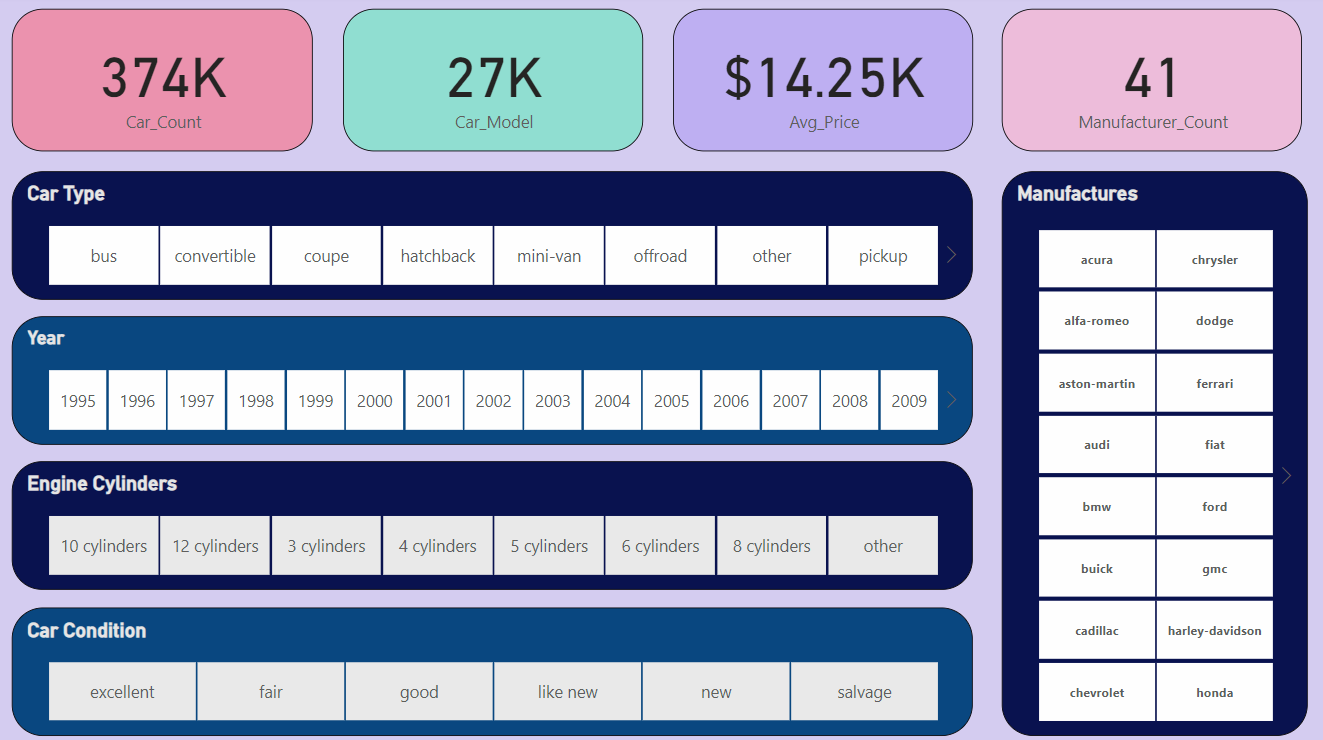
**Fig.: Bar chart of Count of Model Vs. Manufacturer**

****

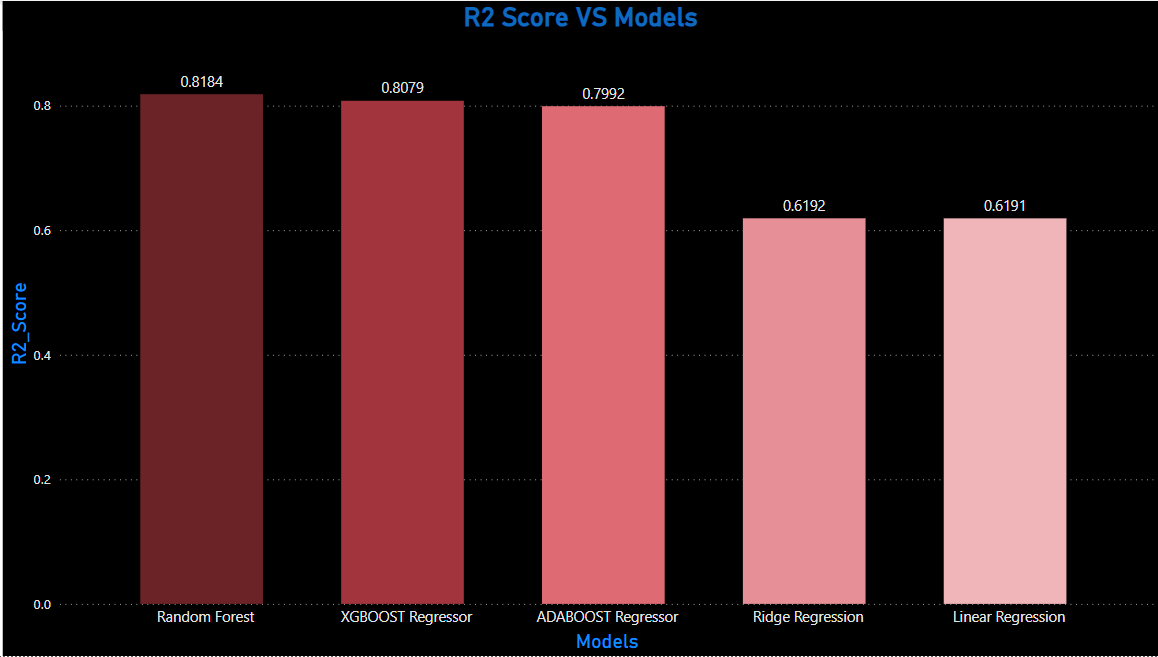
**Fig.: Bar chart of Car Price Vs. Manufacturer.**

****

**Fig.: Bar Chart of Average Price Vs. Fuel and Car Condition.**

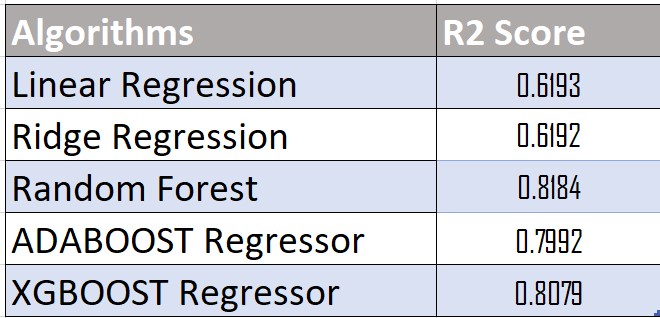


**Fig.: Dashboard by filters.**

****

**Fig.: Overall Model Performance (R2 score Vs. Model)**

# Result and Finding



# 8. CONCLUSION AND FUTURE SCOPE

The increased prices of new cars and the financial incapability of the customers to buy them, Used Car sales are on a global increase. Therefore, there is an urgent need for a Used Car Price Prediction system which effectively determines the worthiness of the car using a variety of features. The proposed system will help to determine the accurate price of used car price prediction.

In future this machine learning model may bind with various website which can provide real time data for price prediction. Also, we may add large historical data of car price which can help to improve accuracy of the machine learning model. We can build an android app as user interface for interacting with user. For better performance, we plan to judiciously design deep learning network structures, use adaptive learning rates and train on clusters of data rather than the whole dataset.

# REFERENCES

* <https://github.com/abhashpanwar/used-car-price-prediction>.