A PROJECT ON

"USED CAR PRICE PREDICTION"

SUBMITTED IN
PARTIAL FULFILLMENT OF THE
REQUIREMENT FOR THE COURSE OF
DIPLOMA IN BIG DATA ANALYSIS

SUNBEAM INSTITUTE OF INFORMATION TECHNOLOGY, PUNE

Submitted By:

Abhsihek Singh (80490)

Rahul Chaurasiya (80476)

Mr.Nitin Kudale
Centre Coordinator

Mrs.Manisha Hingne
Course Coordinator

CERTIFICATE

This is to certify that the project work under the title ‘Used Car Price Prediction’ is done by Abhishek Singh & Rahul Chaurasiya in partial fulfillment of the requirement for award of Diploma in Big Data Analysis Course.

Mr. Aniket P
Project Guide

Mrs. Manisha Hingne
Course Coordinator

Date:22/02/2024

ACKNOWLEDGEMENT

A project usually falls short of its expectation unless aided and guided by the right persons at the right time. We avail this opportunity to express our deep sense of gratitude towards Mr. Nitin Kudale (Center Coordinator, SIIT, Pune) and Mrs. Manisha Hingne (Course Coordinator, SIIT ,Pune) and Project Guide Mr. Aniket P.

We are deeply indebted and grateful to them for their guidance, encouragement and deep concern for our project. Without their critical evaluation and suggestions at every stage of the project, this project could never have reached its present form.

Last but not the least we thank the entire faculty and the staff members of Sunbeam Institute of Information Technology, Pune for their support.

Abhishek Singh
DBDA September 2023
Batch, SIIT Pune

Rahul Chaurasiya
DBDA September 2023
Batch, SIIT Pune

.

TABLE OF CONTENTS

1. Introduction

1.1 Introduction And Objectives:

Introduction:

Our project focuses on leveraging machine learning techniques to predict the prices of second-hand cars. With the burgeoning market for pre-owned vehicles, accurate pricing is crucial for buyers and sellers alike. By harnessing historical data and advanced algorithms, we aim to develop a reliable model that assists in determining fair market values, enhancing transparency and efficiency in the automotive resale sector.

Objective:

The primary objective of our project is to create a machine learning model capable of accurately predicting the prices of second-hand cars. By analyzing various features such as make, model, year, mileage, and condition, our goal is to provide buyers and sellers with valuable insights into fair market values. Ultimately, we aim to streamline the process of buying and selling used vehicles, facilitating informed decision-making and ensuring fair transactions.

1.2 Why this problem needs To be Solved?

The creation of our second-hand car price prediction project is imperative due to several factors. Firstly, in an increasingly dynamic automotive market, the need for accurate pricing of pre-owned vehicles has become paramount. Buyers seek transparency and assurance of fair value, while sellers aim to optimize returns on their investments. Secondly, the sheer volume of data available from past transactions presents an opportunity to leverage machine learning for predictive analysis. Thirdly, by developing a reliable pricing model, we can mitigate the risks associated with overpaying or underselling, fostering trust and efficiency in the resale ecosystem. Ultimately, this project addresses a critical need in the automotive industry, enhancing market dynamics and empowering consumers with valuable insights.

1.3 Dataset information - Vehicles.csv - It is file which contains more than 21 columns and contains information on which website was used car registered for resale , what brand was it , which model , along with engine type , fuel type and other relevant data

1.4 Problem Definition and Algorithm:

a. Problem Definition

The objective of this project is to predict the selling price of used cars based on various features such as make, model, year, mileage, and condition. We are provided with a dataset containing information about used cars, including their selling prices. Our task is to build a model using machine learning algorithms that can accurately predict the selling price of used cars. The dataset is divided into a training set and a test set, and our goal is to fit the model to the training data and evaluate its performance on the test data. Our primary metrics of interest will be the Mean Absolute Error and R2 score, which will measure the accuracy of our predictions. Our goal is to minimize the difference between the predicted selling prices and the actual selling prices of used cars. While it's unlikely to achieve perfect predictions, our aim is to develop a model that can provide as close an estimate as possible to the actual selling price.

b. Algorithim used -

1. **Linear Regression**: A basic regression algorithm that models the relationship between the independent variables (car features) and the dependent variable (car price) by fitting a linear equation to the observed data.

2. **Decision Trees:** Decision trees recursively split the data into subsets based on the most significant attribute. They are intuitive, easy to interpret, and can capture non-linear relationships between features and the target variable.

3. **Random Forest:** A popular ensemble learning technique that combines multiple decision trees to improve performance and reduce overfitting. It works by building multiple decision trees and averaging their predictions.

4. **Gradient Boosting Trees:** Another ensemble method that builds trees sequentially, where each tree corrects the errors made by the previous one. Gradient boosting algorithms like XGBoost, LightGBM, and CatBoost are widely used for regression tasks due to their high performance.

5. **Support Vector Machines (SVM):** SVMs are versatile algorithms that can be used for both classification and regression tasks. In regression, SVMs aim to find the hyperplane that best separates the data points while minimizing the error.

6. **Neural Networks:** Deep learning models, such as feedforward neural networks or more complex architectures like convolutional neural networks (CNNs) or recurrent neural networks (RNNs), can be used for regression tasks. They can capture complex patterns in the data but may require more data and computational resources.

7. **K-Nearest Neighbors (KNN):** KNN is a non-parametric algorithm that predicts the value of a data point by averaging the values of its k nearest neighbors. It's simple to implement but may suffer from the curse of dimensionality.

8. **Ensemble Methods:** Apart from random forests and gradient boosting, other ensemble techniques like bagging and stacking can be used to combine the predictions of multiple base learners to improve accuracy.

2.Experimental Evaluation:

2.1 Methodology:

The objective of this project is to predict the prices of used cars . The data set is contained from Kaggle and has 1 csv file known as vehicles.Loading in raw data

```
df = pd.read_csv("vehicles.csv")
```

Preprocessing:

The car details are given for within the range of Years 1962-20120 for each car. This data was appropriately converted from one data type to other and row values

```
df1['year']=df1['year'].astype(object)
df1 = df1[df1['cylinders']!="other"]
df1['odometer']=df1['odometer'].astype(int)
df1 = df1[df1['type']!="other"]
df1 = df1[df1['price']!="other"]
df2['year']=df2['year'].astype(int)
```

The data had several missing values and needed to be cleaned. Since the number of missing values were significant, they were droped.

Various values were level endoded to facilitate further operations on them

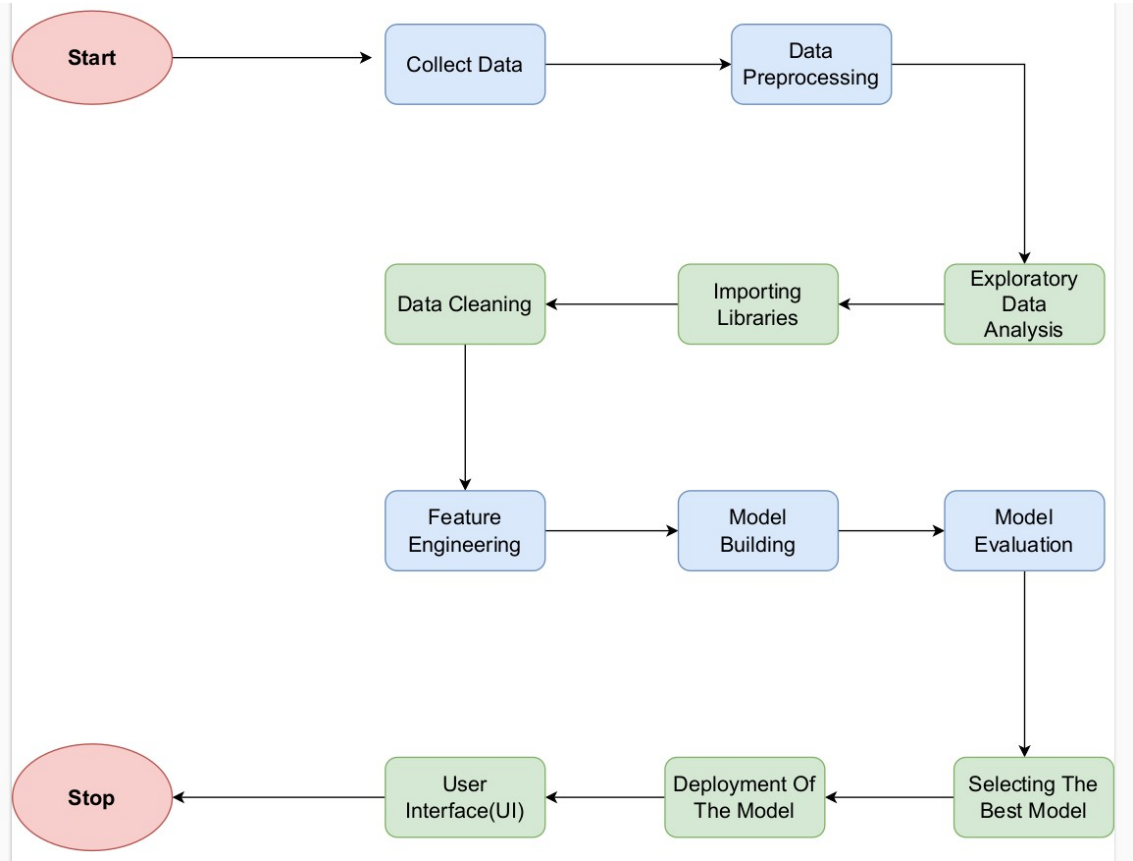
```
encoder = LabelEncoder()
encoder.fit(df2["manufacturer"])
df2["manufacturer"]
=encoder.transform(df2["manufacturer"]
)

encoder = LabelEncoder()
encoder.fit(df2["model"])
```

```
df2["model"]
=encoder.transform(df2["model"])

encoder = LabelEncoder()
encoder.fit(df2["condition"])
df2["condition"]
=encoder.transform(df2["condition"])

encoder = LabelEncoder()
encoder.fit(df2["fuel"])
df2["fuel"]
=encoder.transform(df2["fuel"])

encoder = LabelEncoder()
encoder.fit(df2["title_status"])
df2["title_status"]
=encoder.transform(df2["title_status"])


encoder = LabelEncoder()
encoder.fit(df2["transmission"])
df2["transmission"]
=encoder.transform(df2["transmission"]
)


encoder = LabelEncoder()
encoder.fit(df2["drive"])
df2["drive"]
=encoder.transform(df2["drive"])


encoder = LabelEncoder()
encoder.fit(df2["type"])
df2["type"]
=encoder.transform(df2["type"])
```

```
encoder = LabelEncoder()
encoder.fit(df2["paint_color"])
df2["paint_color"]
=encoder.transform(df2["paint_color"])
```

```
encoder = LabelEncoder()
encoder.fit(df2["state"])
df2["state"]
=encoder.transform(df2["state"])
```

Project flow diagram:



3.2 Exploratory Data Analysis

The total market size for the all the used cars sold in each state is given in the form of map chart , where the darker colour shows higher capitalization . From here we can observe california has the highest market capitalization of all the states present.

Fig 1: Map-chart showing market capitalization by state

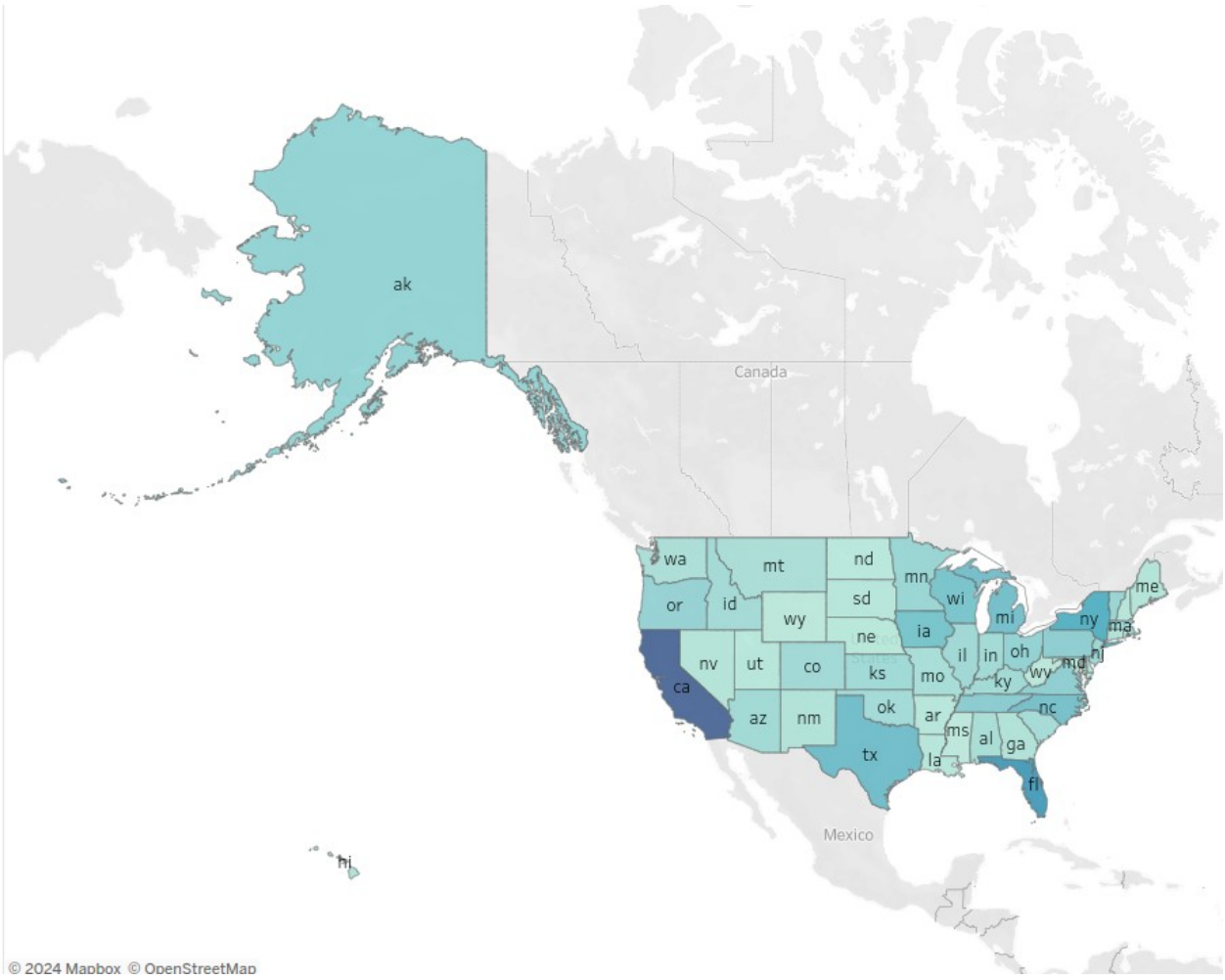
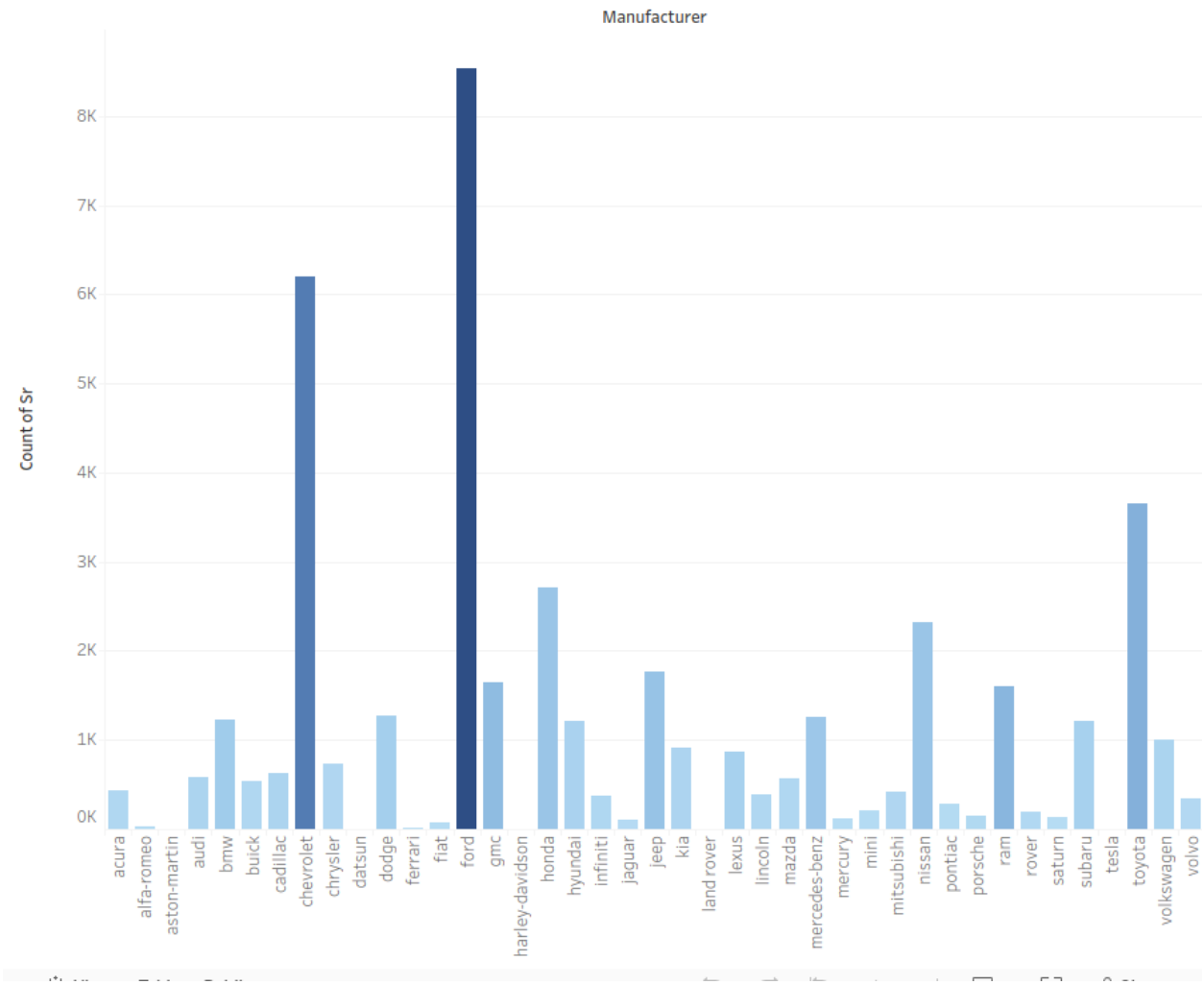
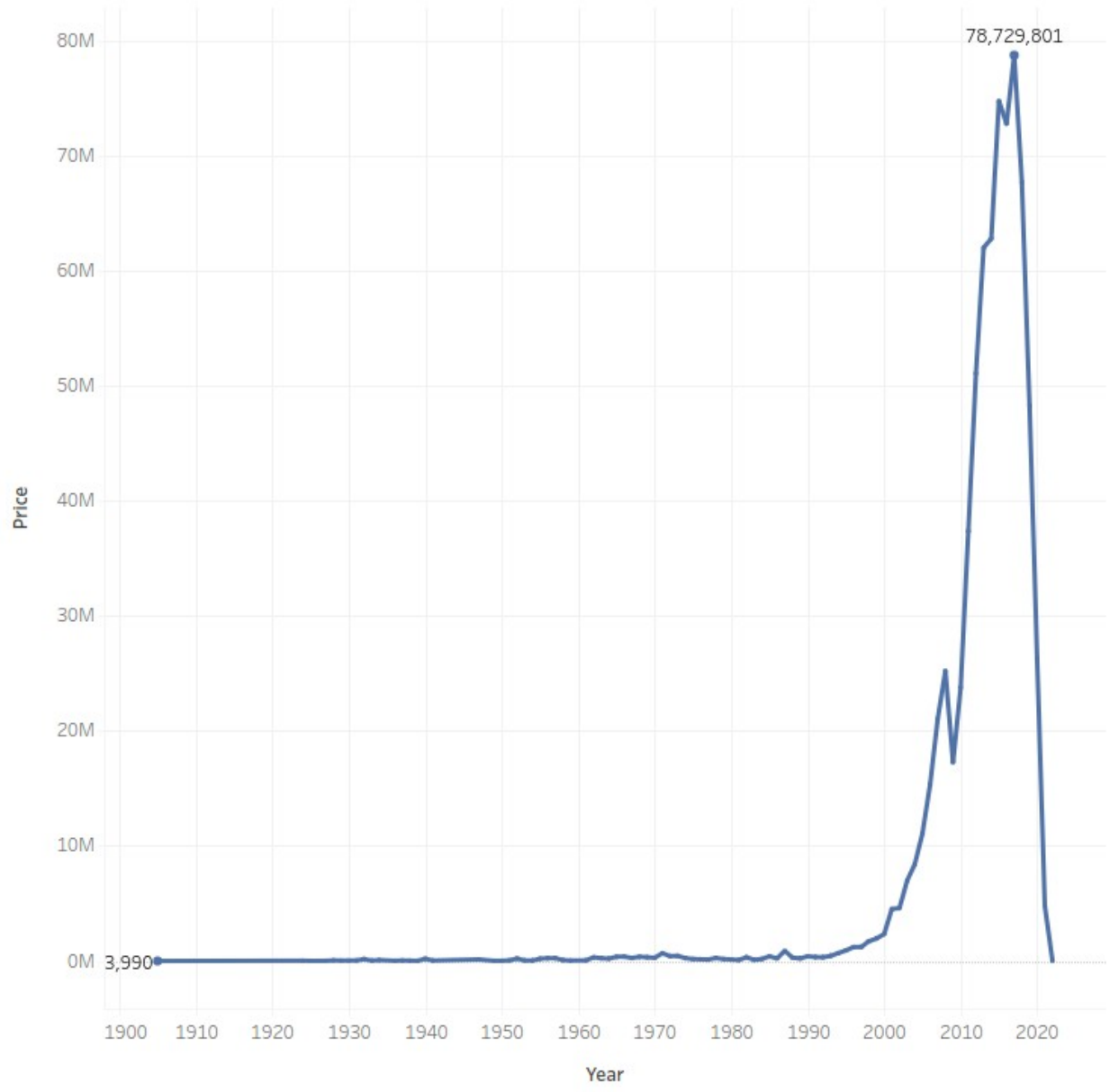


Fig 2: Bar-chart showing no of used cars sold by company



from bar chart shown we can infer that while no of used cars sold by ford company were the highest in number while , they were least in case of tesla

Fig 3:Line-plot:Shows year wise total price of the used cars sold



From this graph we can infer that total price of the used car sold was highest for the year 2017 , and it was lowest for the year 1905

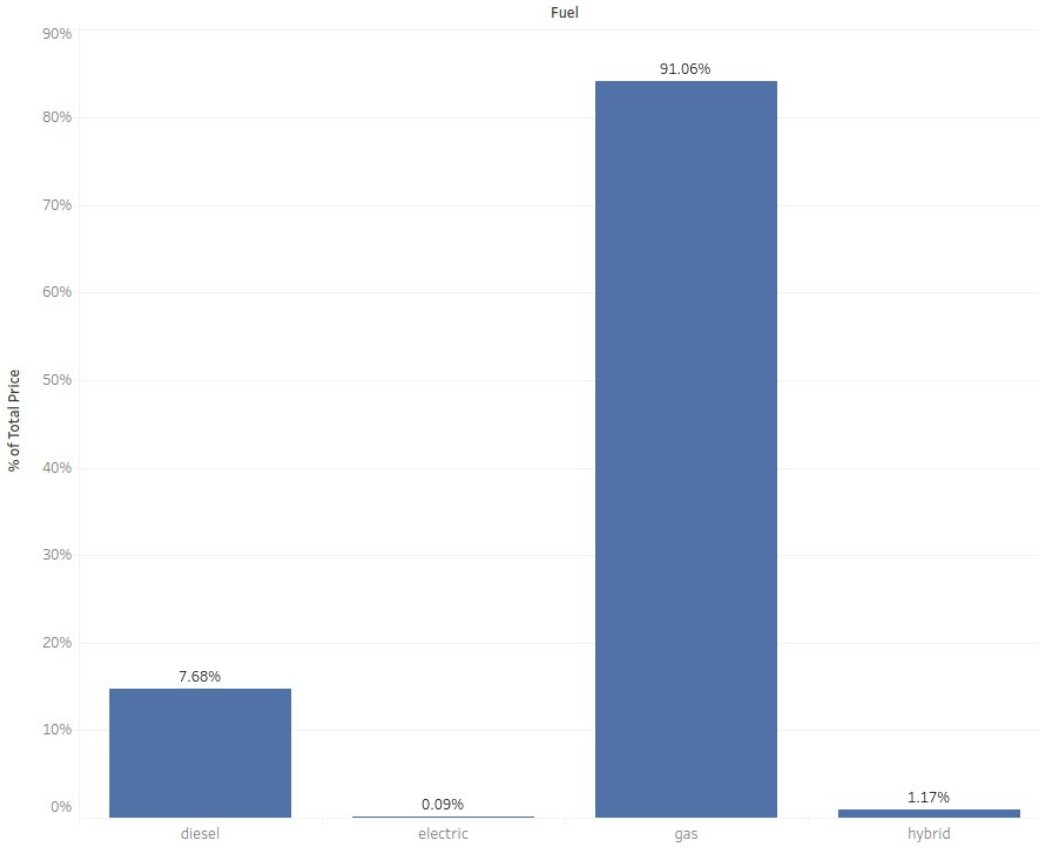


Fig 4:Bar chart- Showing the percentage distribution of type of car-sold by fuel type

The bar chart shows that most number of used cars sold were gas-type cars , and least no of cars sold were electric-type.

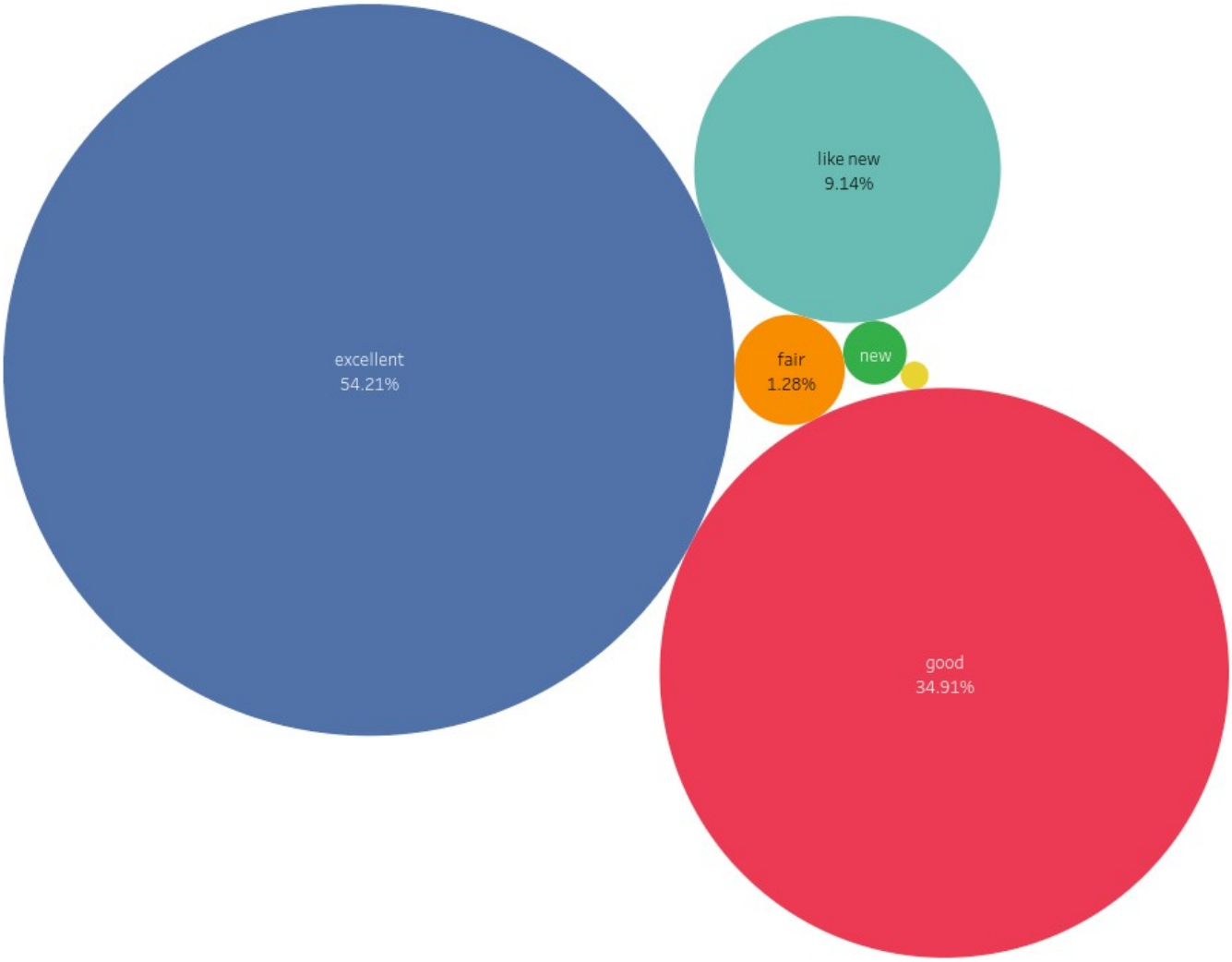


Fig 6: Bubble-chart:Showing the percentage distribution of cars sold by condition-label.

The used cars sold were given label depicting their condition. The labels were “fair” , “good” , “new”, “like new”,”salvage”,and “excellent”.It was found that most used cars sold were labelled as “excellent” , and least number of cars were “salvage” condition.

3. Results and discussion:

Decision-Tree, linear regression , logistic regression, k-nearest neighbours(knn) , random forest, decision tree , xgboost , catboost machine algorithm were used to predict the prices of the used cars. Among the given algorithms Random forsest Machine algorithm was the best performing one as it provided the highest R2 score of 0.86

```
  def build_model_RF():
from sklearn.ensemble import RandomForestRegressor

model = RandomForestRegressor()

model.fit(x_train,y_train)

# with open('RandomForest001.pkl', 'wb') as file:
#     pickle.dump(model, file)


return model
```

MAE : 2218
MSE:2069
R2 Score : 0.87

4. GUI:

GUI is made using Flask framework. **Flask** is a micro web framework written in Python. It is classified as a microframework because it does not require particular tools or libraries. It has no database abstraction layer, form validation, or any other components where pre-existing third-party libraries provide common functions. However, Flask supports extensions that can add application features as if they were implemented in Flask itself. Extensions exist for object-relational mappers, form validation, upload handling, various open authentication technologies and several common framework related tools

6.GitHubLink: https://github.com/b180490/Project1.git

7.Future work And Conclusion

The creation of our second-hand car price prediction project is imperative due to several factors. Firstly, in an increasingly dynamic automotive market, the need for accurate pricing of pre-owned vehicles has become paramount. Buyers seek transparency and assurance of fair value, while sellers aim to optimize returns on their investments. Secondly, the sheer volume of data available from past transactions presents an opportunity to leverage machine learning for predictive analysis. Thirdly, by developing a reliable pricing model, we can mitigate the risks associated with overpaying or underselling, fostering trust and efficiency in the resale ecosystem. Ultimately, this project addresses a critical need in the automotive industry, enhancing market dynamics and empowering consumers with valuable insights.