ABSTRACT

The project titled "Car Price Prediction" addresses the crucial task of predicting car prices using machine learning techniques. In an ever-evolving automotive market, accurate car price predictions are significant for both buyers and sellers. This project employs a dataset of 8,200 car listings to develop and evaluate predictive models for car prices.

The methodology involves data collection, preprocessing, and feature engineering. Machine learning models, including Random Forest, Gradient Boosting, and Extra Trees regressors, are applied to the prepared dataset. Key evaluation metrics such as R-squared (R2), Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) are used to assess the model performance.

The results demonstrate the effectiveness of the machine learning models in predicting car prices, with model comparison revealing their respective strengths and weaknesses. Accurate car price predictions can significantly benefit car buyers, sellers, and market analysts, aiding in informed decision-making and market insights.

This project contributes to the field of predictive modeling in the automotive industry, offering valuable insights into the application of machine learning for car price predictions.

TABLE OF CONTENT

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 GENERAL

1.1.1 PROJECT OVERVIEW

Car prices are a subject of immense importance and relevance in the ever-evolving automotive market. As one of the most dynamic sectors in the global economy, the automotive industry constantly adapts to shifting consumer preferences, market trends, and technological advancements. The challenge of accurately determining car prices amidst this dynamism is a crucial one that affects both buyers and sellers.

1.1.1.1 INTRODUCTION TO CAR PRICE PREDICTION

The core objective of this project is to explore the application of machine learning techniques to predict used car prices. Used Car price prediction involves estimating the fair value of a vehicle, which depends on various factors, such as the make and model, year, mileage, condition, location, odometer measure, etc,.. Traditional pricing methods often struggle to comprehensively consider these variables, leading to inaccurate valuations.

1.1.1.2 ROLE OF DATA SCIENCE AND MACHINE LEARNING

Data science and machine learning offer a data-driven approach to solving complex problems, making them well-suited to address the challenges in car price prediction. By analyzing historical car listings data, we can develop predictive models that take into account a wide range of features and factors that influence car prices. The ability to accurately predict used car prices has far-reaching implications, from helping buyers make informed decisions to enabling sellers to set competitive prices.

This project explores the potential of data science and machine learning to improve the accuracy and reliability of car price predictions. By leveraging a dataset of 8,200 car listings, we aim to develop models that provide valuable insights into the fair market values of different vehicles. The results of this project can contribute to the efficiency of the automotive market and empower both buyers and sellers with data-driven pricing strategies.

1.1.2 DATA SOURCE

The foundation of this project's research is the used car listings dataset, which is integral to the process of predicting car prices accurately. This dataset comprises detailed records of 8,200 car listings, encompassing a diverse range of vehicles available for sale in the automotive market. The dataset includes comprehensive information for each listing, allowing for a comprehensive analysis of the factors influencing car prices.

1.1.2.1 DESCRIPTION OF CAR LISTINGS DATASET

The cornerstone of our project is the car listings dataset, which serves as the primary source of information for predicting car prices. This dataset comprises records of 8,200 car listings, each offering a comprehensive profile of vehicles available for sale. The dataset includes essential information such as the model, year of manufacture, mileage, condition, location, price, and various features.

This dataset offers a diverse collection of used car listings, encompassing a wide range of car models. It provides insight into the variations in car prices across different vehicle types, brands, and geographical regions. Each listing is accompanied by detailed information that allows for a comprehensive analysis of the factors influencing car prices.

1.1.2.2 DATA RELEVANCE AND QUALITY

The relevance and quality of the dataset are paramount in ensuring the accuracy of our predictive models. Rigorous efforts were made to curate a dataset that is not only comprehensive but also of high quality. This involved data cleansing processes to remove any inconsistencies, inaccuracies, or outliers that could distort the predictive models.

Data relevance was assessed by considering factors such as the recency of listings, the completeness of information, and the representation of different car types. Data quality was ensured through extensive validation, which included cross-referencing listing information with industry standards and market trends.

In our subsequent chapters, we will delve deeper into the specifics of data collection and preprocessing, where the steps taken to ensure data relevance and quality will be elucidated.

1.2 BACKGROUND

The field of predictive modeling and machine learning has gained significant momentum in recent years. In various domains, including finance, healthcare, and e-commerce, machine learning models have demonstrated their ability to make accurate predictions and inform decision-making processes. The automotive industry can also benefit from these techniques, particularly in the context of predicting car prices.

Car prices are influenced by a multitude of factors, such as the vehicle's model, year, mileage, condition, location, and various features. Traditional pricing methods often fall short in comprehensively considering these factors, resulting in inaccurate valuations. Machine learning provides a data-driven approach that can take into account these diverse variables and improve the precision of car price predictions.

1.2.1 CHALLENGES IN CAR PRICING

The pricing of cars is a multifaceted process influenced by various factors, and it presents a unique set of challenges. Understanding and addressing these challenges is crucial for the development of accurate car price prediction models. Some of the key challenges in car pricing include:

1.2.1.1 FACTORS AFFECTING CAR PRICES

Car pricing is a multifaceted process influenced by an array of factors. These factors can vary significantly between vehicles and markets. Some of the primary factors affecting car prices include:

Vehicle Model: The brand and model of a car often play a substantial role in determining its price. Well-established and prestigious brands tend to command higher prices.

Year of Manufacture: The age of a car, as indicated by its manufacturing year, has a direct impact on pricing. Newer cars are generally priced higher than older ones.

Mileage: The number of miles a car has traveled is a key determinant. Lower mileage is typically associated with higher prices.

Condition: The overall condition of the vehicle, including any wear and tear or damage, significantly influences its value. New or well-maintained cars are valued higher.

Location: Geographical location is another important factor. The same car may be priced differently in various regions due to variations in supply and demand.

Features and Specifications: Specific features and specifications of the car, such as its engine type, transmission, interior materials, and additional accessories, can also impact pricing.

Understanding how these factors interplay in the determination of car prices is essential for our project. By leveraging machine learning techniques, we aim to capture and incorporate these variables in our predictive models.

1.2.1.2 MARKET VOLATILITY AND DYNAMIC PRICING

The automotive market is known for its volatility. Car prices can fluctuate rapidly in response to changes in supply and demand, economic conditions, and external events. This market volatility poses challenges for both buyers and sellers. In dynamic pricing scenarios, traditional valuation methods often fall short, as they are unable to adapt to rapidly changing conditions.

Dynamic pricing strategies, often employed by sellers, involve adjusting car prices in real-time to maximize revenue. This flexibility is driven by data analysis and market trends, making accurate price predictions an imperative aspect of competitiveness. Machine learning models can offer a robust solution to address the complexities of dynamic pricing, ensuring that vehicles are competitively priced in response to market fluctuations.

1.2.1.3 TRADITIONAL VALUATION METHODS

Traditional methods of car valuation often rely on historical pricing data and expert judgment. These methods, while valuable, are limited in their ability to comprehensively consider all the factors that influence car prices. They may not adapt well to market dynamics and can result in valuations that deviate significantly from actual market values.

Inaccurate valuations can lead to various issues, including overpricing, which may deter potential buyers, or underpricing, which can result in financial losses for sellers. Our project aims to bridge these gaps by utilizing data-driven machine learning models that can consider a wider array of variables, thus providing more precise and adaptable car price predictions.

1.2.2 IMPORTANCE OF ACCURATE CAR PRICE PREDICTIONS

Accurate car price predictions hold significant importance for various stakeholders in the automotive market. Understanding the value of precise pricing is crucial to appreciating the impact of our project. The following key points underscore the significance of accurate car price predictions.

1.2.2.1 BUYER'S CHALLENGE IN PRICING

For prospective buyers, accurate car price predictions provide valuable guidance. Buyers can make informed decisions, confidently assessing whether a listed price aligns with the fair market value of a vehicle. This knowledge empowers buyers to negotiate effectively and make sound financial choices.

1.2.2.2 SELLER'S NEED FOR COMPETITIVE PRICING

Sellers, on the other hand, require competitive pricing strategies to attract potential buyers. Overpricing may result in listings going unnoticed, while underpricing can lead to financial losses. Accurate predictions enable sellers to set competitive prices, maximizing the likelihood of selling their vehicles swiftly and profitably.

1.2.2.3 MARKET EFFICIENCY AND INFORMED DECISION-MAKING

Beyond individual benefits, accurate car price predictions contribute to the overall efficiency of the automotive market. When buyers and sellers have access to reliable pricing information, market transactions become more efficient, and the allocation of resources is optimized. Informed decision-making prevails, promoting transparency and trust within the automotive market.

1.3 OBJECTIVES

The primary objectives of this project are as follows:

- To collect and preprocess a comprehensive dataset of car listings, ensuring data quality and relevance.

- To apply machine learning techniques, including Random Forest, Gradient Boosting, and Extra Trees regressors, to develop predictive models for car prices.
- To evaluate and compare the performance of these models using key metrics such as R-squared (R2), Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE).
- To provide valuable insights into the effectiveness of Data Science in the context of car price prediction.
- To contribute to the automotive industry by offering a data-driven approach that enhances the accuracy of car price predictions, benefiting both buyers and sellers.

In the subsequent chapters, we will delve into the methodology, results, and discussions surrounding this project, culminating in a comprehensive understanding of the application of machine learning in predicting car prices.

CHAPTER 2

LITERATURE REVIEW

2.1 GENERAL

In this chapter, we delve into the existing body of knowledge surrounding car price prediction and machine learning techniques. A comprehensive literature review serves as the foundation for understanding the current state of the field and the gaps that our project aims to address.

2.2 MACHINE LEARNING IN PREDICTIVE MODELING

Machine learning stands at the forefront of predictive modeling, driving advancements in various industries, including the automotive sector. In the context of car price prediction and other automotive applications, machine learning algorithms have proven invaluable. This section explores the pivotal role of machine learning in predictive modeling within the automotive domain.

2.2.1 MACHINE LEARNING APPLICATIONS IN THE AUTOMOTIVE SECTOR

Machine learning techniques have revolutionized predictive modeling in the automotive sector. By leveraging vast datasets and advanced algorithms, these methods have enabled more accurate and adaptable pricing models. We explore the specific applications of machine learning in the automotive domain, including the development of pricing models and their real-world utility.

2.2.1.1 CAR PRICE PREDICTION MODELS

Car price prediction models represent a pivotal application of machine learning in the automotive sector. These models leverage advanced algorithms and large datasets to estimate the market value of vehicles accurately. By considering a multitude of variables such as the make, model, age, mileage, condition, and location, these models provide valuable insights into vehicle pricing.

2.2.1.2 DEMAND FORECASTING

Demand forecasting is a crucial facet of machine learning's role in the automotive sector. It involves the prediction of market trends, consumer preferences, and the demand for specific vehicle models. These forecasts enable automakers, dealerships, and suppliers to make informed decisions regarding vehicle production and distribution, thereby enhancing efficiency and profitability.

2.2.1.3 VEHICLE HEALTH MONITORING

Machine learning plays a transformative role in monitoring the health of vehicles. Predictive maintenance systems rely on machine learning to assess a vehicle's condition continuously. By analyzing sensor data, usage patterns, and historical maintenance records, these systems predict when maintenance or repairs are needed, reducing downtime and ensuring vehicle longevity.

2.2.1.4 AUTONOMOUS VEHICLES

The development and deployment of autonomous vehicles mark a cutting-edge application of machine learning. These vehicles employ sensors, cameras, and deep learning algorithms to perceive their surroundings, interpret road conditions, and make real-time decisions. Machine learning is at the core of achieving safe and reliable autonomous driving, promising a future of safer and more efficient transportation.

2.2.2 MODEL TYPES AND ALGORITHMS

Machine learning offers a diverse array of model types and algorithms for predictive modeling. This section examines the various models employed in car price prediction, including regression models, decision trees, ensemble methods, and neural networks. Each model type's strengths and weaknesses in the context of car price prediction are evaluated.

2.2.2.1 REGRESSION-BASED MODELS

Regression-based models serve as a fundamental category of predictive modeling techniques within the automotive sector. These models focus on establishing a relationship between various vehicle features and the final sale price.

2.2.2.2 DECISION TREES AND ENSEMBLE METHODS

In addition to regression-based models, decision trees and ensemble methods offer powerful alternatives for car price prediction:

Decision Trees: Decision tree models are non-linear and non-parametric. They break down the prediction process into a series of binary decisions based on feature values. Decision trees can handle both numerical and categorical features, making them adaptable to diverse datasets.

Random Forests: Random Forests are ensemble methods that combine multiple decision trees to improve prediction accuracy and reduce overfitting. By aggregating the predictions of individual trees, Random Forests provide robust and versatile solutions for car price prediction.

Gradient Boosting: Gradient Boosting is another ensemble technique that sequentially builds a series of decision trees, where each tree corrects the errors of the previous one. This method often leads to exceptionally accurate predictive models, making it a valuable tool in car price prediction scenarios.

2.2.3 MODEL EVALUATION METRICS

An essential aspect of machine learning in predictive modeling is the choice of appropriate evaluation metrics. We explore the metrics commonly used to assess model performance, such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R2). Understanding these metrics is critical for comparing and selecting the best-performing models.

2.2.3.1 MEAN SQUARED ERROR (MSE)

The Mean Squared Error (MSE) is a metric used to evaluate the performance of predictive models in the automotive sector. It measures the average of the squared differences between the predicted car prices and the actual car prices. A lower MSE indicates that the model's predictions are closer to the actual prices, reflecting better accuracy.

2.2.3.2 ROOT MEAN SQUARED ERROR (RMSE)

The Root Mean Squared Error (RMSE) is a variation of the MSE and is widely used for assessing the accuracy of car price prediction models. RMSE represents the square root of the MSE, which results in a metric with the same unit as the car prices. A lower RMSE signifies that the model's predictions are closer to the actual prices, making it easier to interpret the level of error in the original price units.

2.2.3.3 R-SQUARED (R2) SCORE

The R-squared (R2) score, often referred to as the coefficient of determination, is a valuable metric for assessing the goodness of fit of predictive models in the automotive sector. It quantifies the proportion of the variance in car prices that can be explained by the model. An R2 score of 1 indicates that the model perfectly predicts car prices, while a score of 0 suggests that the model does not explain any variance. Higher R2 scores are indicative of better model performance in explaining price variations.

2.3 PREVIOUS CAR PRICE PREDICTION STUDIES

In recent years, the automotive industry has witnessed a growing body of research and studies focused on car price prediction. These studies have been instrumental in harnessing data-driven solutions to better understand and predict the market value of vehicles. Researchers have explored various data sources, including historical sales data, online listings, and vehicle specifications, to identify the most influential features affecting car prices. Furthermore, a range of predictive models, including regression-based models, decision trees, neural networks, and ensemble methods, has been scrutinized to develop accurate pricing models. The evaluation of these models has typically relied on metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R2) scores to assess their effectiveness.

Another significant focus in previous studies has been on unraveling the dynamics of the automotive market and the pricing strategies employed by dealerships. These studies have considered market volatility, regional variations, seasonality, and other external factors influencing pricing. Understanding consumer behavior, demand forecasting, and preferences has been a central theme, helping automotive stakeholders optimize production and pricing strategies. Additionally, geographic and temporal analyses have revealed insights into regional pricing disparities and the evolution of car prices over time. Collectively, these previous studies have laid the foundation for more precise and data-driven pricing strategies in the automotive sector.

CHAPTER 3

DATA COLLECTION AND PREPROCESSING

3.1 DATA SOURCE

The dataset utilized in this project was graciously provided by OASIS INFOBYTE. The dataset offers a comprehensive collection of car listings, each offering a wealth of information about the vehicles listed for sale.

The data was made available for our project's use, and it covers a broad range of attributes, including vehicle specifications, pricing details, seller information, and additional features.

```
import pandas as pd
data = pd.read_csv("car.csv")
```

3.2 EXPLORATORY DATA ANALYSIS (EDA)

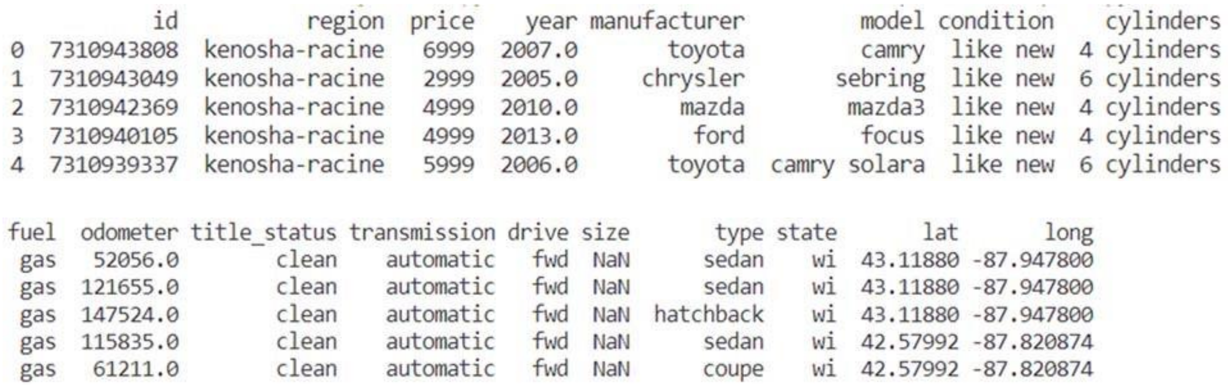
Exploratory Data Analysis (EDA) is a crucial initial step in understanding the dataset and gaining insights into its characteristics. In this section, we present the key visualizations and analyses performed to comprehend the dataset's structure and distribution.

The EDA phase includes the following key components:

3.2.1 DATA SUMMARY

To gain an understanding of the types of vehicles in the dataset, we created a bar chart that illustrates the distribution of different car types. This visualization allows us to see the relative frequencies of various car categories, helping us identify which types are most prevalent.

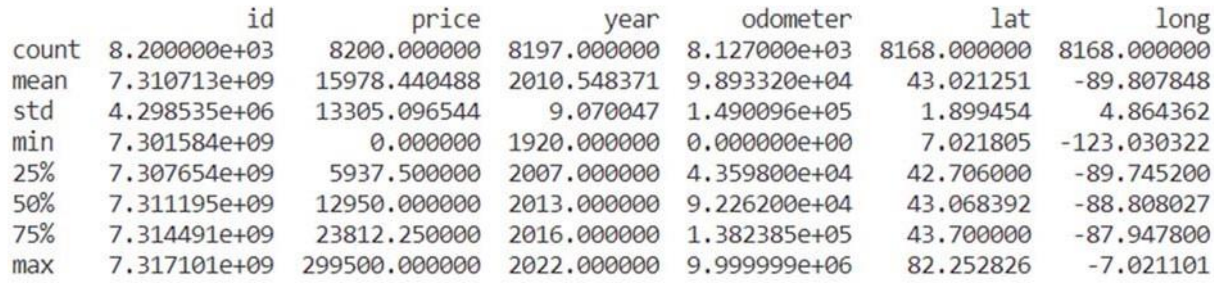
```
print(data.head())
```



```
           id          region  price    year manufacturer         model condition    cylinders
0  7310943808  kenosha-racine   6999  2007.0       toyota         camry  like new  4 cylinders
1  7310943049  kenosha-racine   2999  2005.0     chrysler       sebring  like new  6 cylinders
2  7310942369  kenosha-racine   4999  2010.0        mazda        mazda3  like new  4 cylinders
3  7310940105  kenosha-racine   4999  2013.0         ford         focus  like new  4 cylinders
4  7310939337  kenosha-racine   5999  2006.0       toyota  camry solara  like new  6 cylinders

fuel  odometer title_status transmission drive size       type state       lat       long
 gas   52056.0        clean    automatic   fwd  NaN      sedan    wi  43.11880 -87.947800
 gas  121655.0        clean    automatic   fwd  NaN      sedan    wi  43.11880 -87.947800
 gas  147524.0        clean    automatic   fwd  NaN  hatchback    wi  43.11880 -87.947800
 gas  115835.0        clean    automatic   fwd  NaN      sedan    wi  42.57992 -87.820874
 gas   61211.0        clean    automatic   fwd  NaN      coupe    wi  42.57992 -87.820874
```

```
print(data.describe())
```



```
                 id          price         year      odometer          lat         long
count  8.200000e+03    8200.000000  8197.000000  8.127000e+03  8168.000000  8168.000000
mean   7.310713e+09   15978.440488  2010.548371  9.893320e+04    43.021251   -89.807848
std    4.298535e+06   13305.096544     9.070047  1.490096e+05     1.899454     4.864362
min    7.301584e+09       0.000000  1920.000000  0.000000e+00     7.021805  -123.030322
25%    7.307654e+09    5937.500000  2007.000000  4.359800e+04    42.706000   -89.745200
50%    7.311195e+09   12950.000000  2013.000000  9.226200e+04    43.068392   -88.808027
75%    7.314491e+09   23812.250000  2016.000000  1.382385e+05    43.700000   -87.947800
max    7.317101e+09  299500.000000  2022.000000  9.999999e+06    82.252826    -7.021101
```

3.2.2 FUEL DISTRIBUTION

In this section, we explore the distribution of different fuel types within the dataset. Fuel type is a crucial attribute when considering the environmental and economic implications of a vehicle. Understanding the distribution of fuel types in the dataset provides insights into the predominant choices among vehicle buyers.

```
plt.figure(figsize=(10, 6))
sns.countplot(data=data, y='fuel', palette='viridis')
plt.title('Fuel Distribution')
plt.xlabel('Count')
plt.ylabel('Fuel')
plt.show()
```

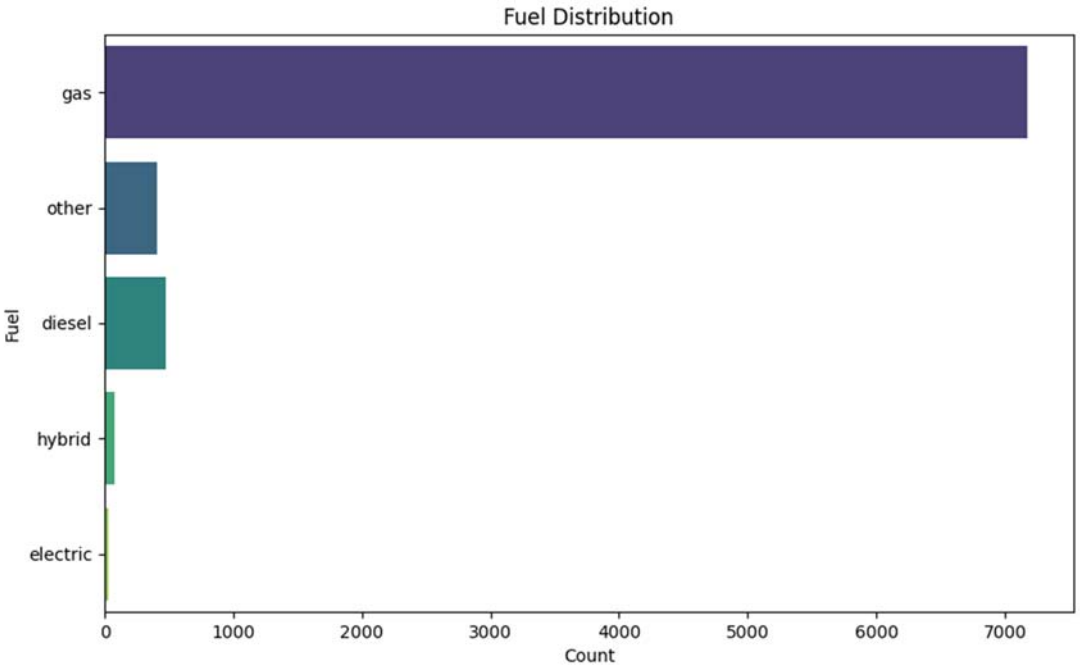


Fig 3.2.2 – Distribution of Fuels

- Gasoline appears to be the most common fuel type among the vehicles in the dataset. It significantly outweighs other fuel types, indicating its popularity in the market.
- Diesel fuel comes second in terms of distribution, though it is notably less common than gasoline-fueled vehicles.
- Other fuel types, including electric and hybrid vehicles, show a smaller but notable presence in the dataset. This reflects the growing interest in more environmentally friendly and fuel-efficient options.

3.2.3 MANUFACTURER DISTRIBUTION

This section explores the distribution of vehicle manufacturers within the dataset. Manufacturers play a significant role in determining the design, quality, and brand identity of a vehicle. Understanding the distribution of manufacturers provides insights into the diversity of brands available in the dataset and potentially reflects market trends.

```
plt.figure(figsize=(40, 30))
sns.countplot(data=data, x='manufacturer', palette='viridis')
plt.title('Bar Plot: Manufacturer Distribution')
plt.ylabel('Count')
plt.xticks(rotation=70, fontsize=6)
plt.show()
```

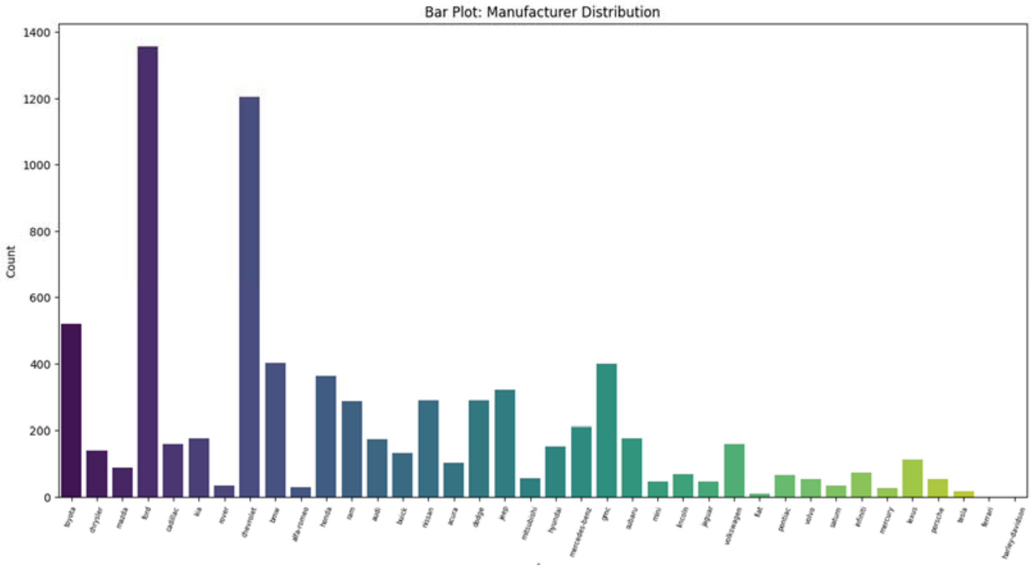


Fig 3.2.3 – Distribution of Manufacturers

3.2.4 CORRELATION ANALYSIS

In this section, we delve into the relationships and correlations between different features or variables within our dataset. A correlation heatmap is a powerful tool that provides a visual representation of how variables relate to one another. By analyzing these correlations, we can gain valuable insights into our data, which can be crucial for making informed decisions in subsequent data analysis or modeling.

```
correlation_matrix = data.corr()
plt.figure(figsize=(20, 15))
sns.heatmap(correlation_matrix, cmap="coolwarm", annot=True)
plt.show()
```

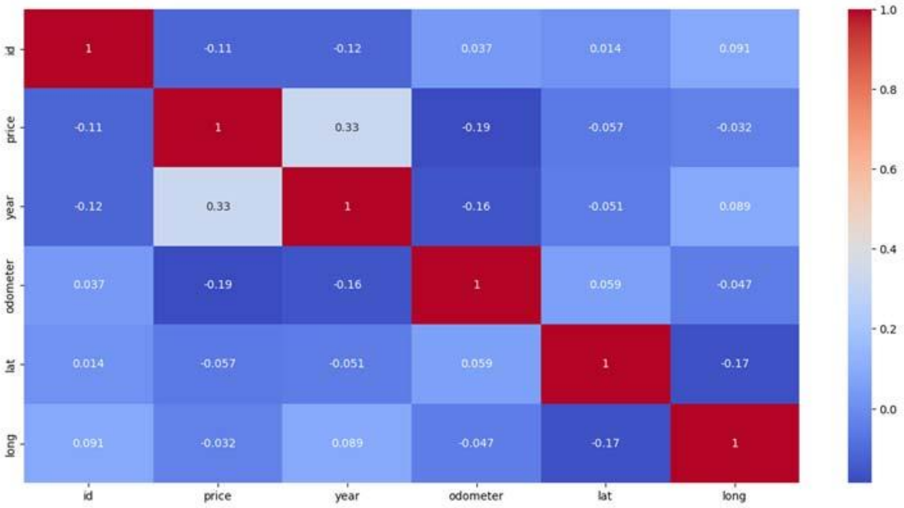


Fig 3.2.4 – Correlation Heatmap

3.3 DATA PREPROCESSING

Data preprocessing is a critical phase in our project, as it ensures the integrity and accuracy of the dataset. To prepare the data for analysis and modeling, we implemented a series of data cleaning procedures:

3.3.1 REMOVING NON-CAR LISTINGS & UNNECESSARY FEATURES

In this step, non-car listings were identified and removed from the dataset based on the 'type' column. The goal of this process was to focus the analysis on genuine car listings and exclude irrelevant entries. Non-car listings, including 'other,' 'van,' and 'bus,' were filtered out to ensure that only valid car data remained for subsequent analysis and modeling.

```
data['type'].unique()
not_cars = data[data['type'].isin(['other', 'van', 'bus'])]
data.drop(not_cars.index, inplace=True)
drop_col = ['lat', 'long']
data = data.drop(columns=drop_col)
```

3.3.2 FILTERING ZERO-VALUED PRICES

To focus on realistic car prices, we filtered out listings with prices less than or equal to $0. This action removes listings with erroneous or non-standard price values.

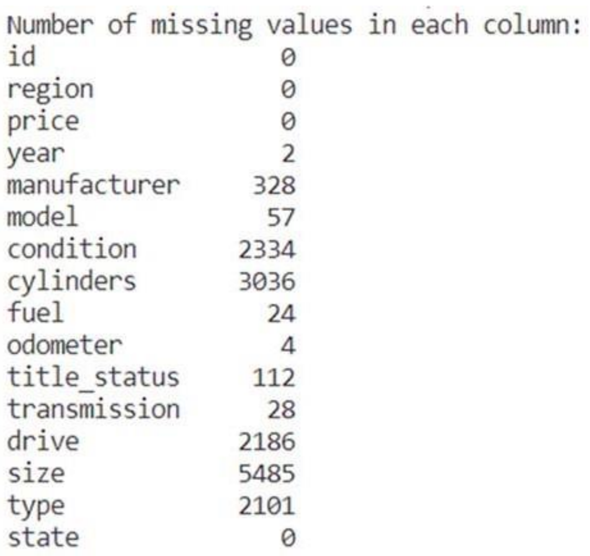
```
zero_price_cars = data[data['price'] == 0]
count_zero_price_cars = zero_price_cars.shape[0]
print("Number of cars listed with a price of '0':", count_zero_price_cars)
data = data[data['price'] > 0]
```

```
Number of cars listed with a price of '0': 181
```

3.3.3 HANDLING MISSING VALUES

We addressed missing data points by removing rows with null values. This step ensures that we work with complete and consistent records, preventing potential issues during analysis.

```
missing_values = data.isna()
num_missing_values = missing_values.sum()
print("Number of missing values in each column:")
print(num_missing_values)
```



```
Number of missing values in each column:
id                 0
region             0
price              0
year               2
manufacturer     328
model             57
condition       2334
cylinders       3036
fuel              24
odometer           4
title_status     112
transmission      28
drive           2186
size            5485
type            2101
state              0
```

```
null_threshold = 1500

columns_to_remove = data.columns[data.isnull().sum() > null_threshold]
data = data.drop(columns=columns_to_remove)
numerical_cols = data.select_dtypes(include=[np.number]).columns
data[numerical_cols] = data[numerical_cols].fillna(data[numerical_cols].mean())
```

3.3.4 OUTLIER DETECTION & HANDLING

We employed an outlier handling strategy, calculated using the Interquartile Range (IQR), to identify and adjust extreme values within the numerical columns. This approach helps mitigate the impact of outliers on the model's accuracy and stability.

```
plt.figure(figsize=(12, 8))
for i, column in enumerate(numerical_cols):
    plt.subplot(2, 2, i + 1)
    sns.boxplot(x=data[column], color='skyblue')
    plt.title(f'Boxplot for {column} (Outliers)')
    plt.xlabel(column)
plt.tight_layout()
plt.show()
```

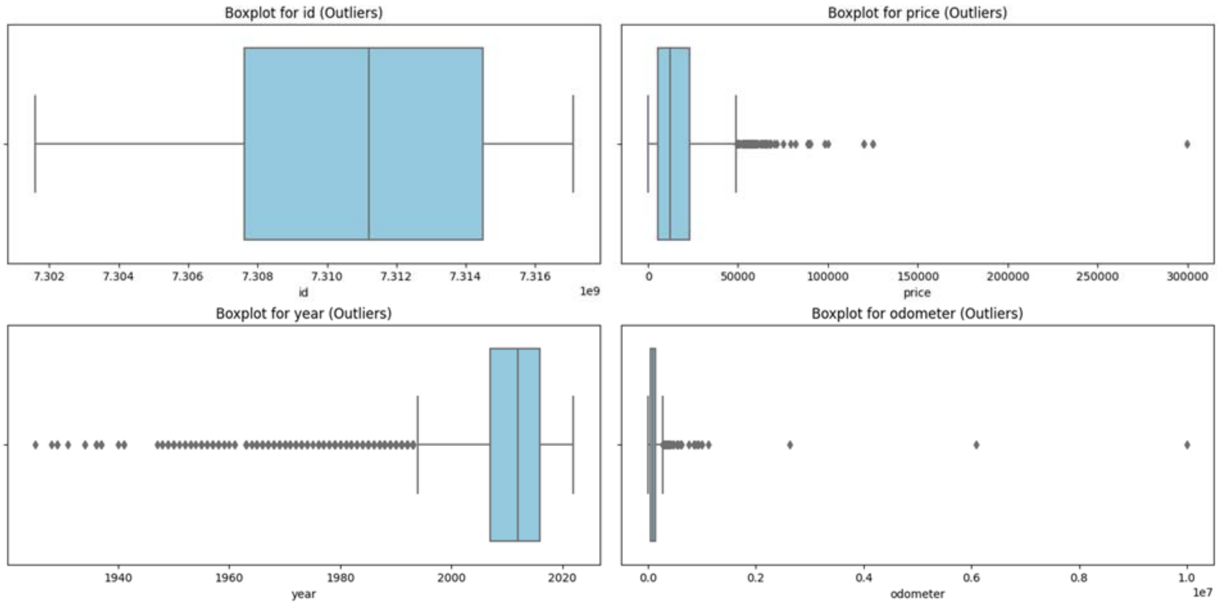


Fig 3.3.4 (a) – Outliers in Numerical Features

```
Q1 = data.quantile(0.25)
Q3 = data.quantile(0.75)
IQR = Q3 - Q1

def handle_outliers(column):
    lower_bound = Q1[column] - 1.5 * IQR[column]
    upper_bound = Q3[column] + 1.5 * IQR[column]
    data[column] = data[column].apply(
        lambda x: max(lower_bound, min(upper_bound, x)))
for column in numerical_cols:
    handle_outliers(column)


plt.figure(figsize=(12, 8))
for i, column in enumerate(numerical_cols):
    plt.subplot(2, 2, i + 1)
    sns.boxplot(x=data[column], color='skyblue')
    plt.title(f'Boxplot for {column} (After Handling Outliers)')
    plt.xlabel(column)
plt.tight_layout()
plt.show()
```

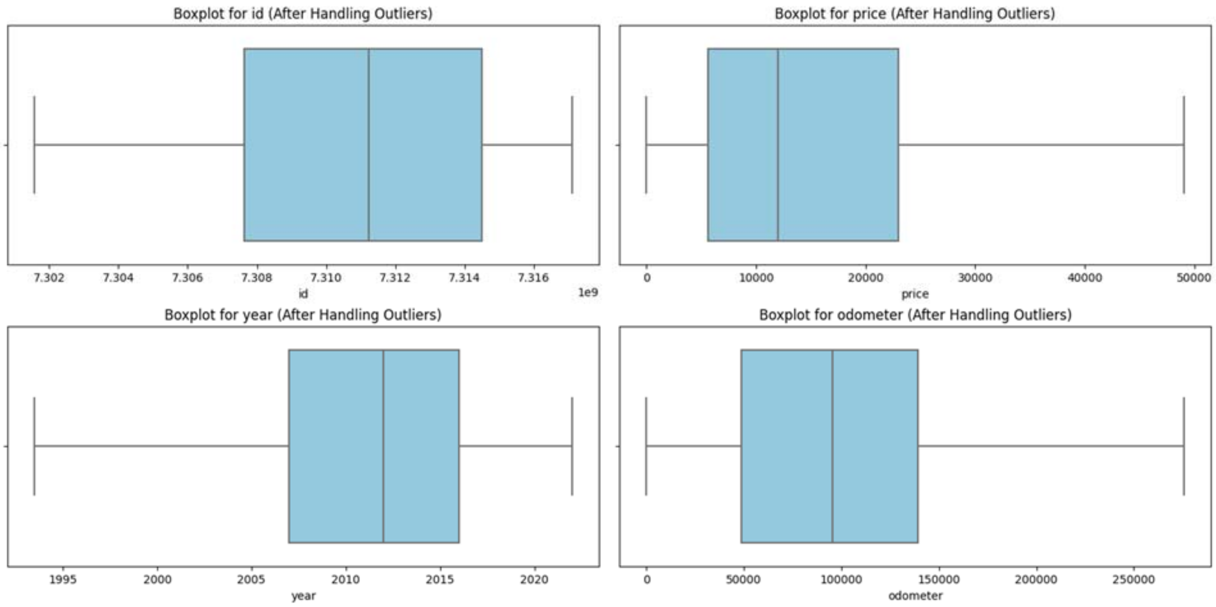


Fig 3.3.4 (b) – Numerical Features after handling Outliers

3.3.5 LABEL ENCODING

To work with categorical features in our machine learning models, we used Label Encoding. This transformation converts non-numeric features into numerical representations, allowing us to incorporate them into the analysis.

```
categorical_cols = data.select_dtypes(exclude=[np.number]).columns
le = LabelEncoder()
for column in categorical_cols:
    data[column] = le.fit_transform(data[column])
```

By executing these data cleaning operations, we have ensured the dataset's consistency and reliability, setting the stage for subsequent analysis and machine learning tasks.

CHAPTER 4

METHODOLOGY

4.1 DATA SPLITTING

Before training a machine learning model, it's essential to split the dataset into two parts: a training set and a testing set. This division allows us to train the model on one portion of the data and assess its performance on another. The process of splitting the data into these subsets is typically performed as follows:

```
selected_features = ['year', 'manufacturer', 'model', 'fuel', 'odometer',
'title_status']
X = data[selected_features]
y = data['price']
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, random_state=32)
```

Once the data is split, we can proceed with training and evaluating our machine learning models on the training and testing sets, respectively. This separation is crucial for assessing the model's generalization performance and detecting issues like overfitting or underfitting.

4.2 MODEL SELECTION

In this section, we'll explore the selection of machine learning models for predicting car prices. The choice of the model plays a crucial role in determining the predictive performance and generalization of the system. We'll consider three different regression models from the scikit-learn library:

```
from sklearn.ensemble import RandomForestRegressor, GradientBoostingRegressor,
ExtraTreesRegressor


rf = RandomForestRegressor(random_state=32)
gb = GradientBoostingRegressor(random_state=32)
et = ExtraTreesRegressor(random_state=32)
```

- **Random Forest Regressor (rf):** The Random Forest Regressor is an ensemble learning method that combines multiple decision tree regressors to improve predictive accuracy and mitigate overfitting.

- **Gradient Boosting Regressor (gb):** The Gradient Boosting Regressor builds an ensemble of decision trees sequentially. It corrects the errors of previous trees and focuses on data points that are harder to predict.

- **Extra Trees Regressor (et):** The Extra Trees Regressor is another ensemble learning method that, like the Random Forest, combines multiple decision trees to make predictions. It adds randomness to the tree-building process, making it less likely to overfit.

4.3 MODEL TRAINING AND PREDICTION

In this section, we will train three different regression models: Random Forest Regressor (rf), Gradient Boosting Regressor (gb), and Extra Trees Regressor (et) using the training dataset. Each of these models is initialized with the same random state for reproducibility.

```
# RandomForestRegressor
rf.fit(X_train, y_train)
rf_y_pred = rf.predict(X_test)


# GradientBoostingRegressor
gb.fit(X_train, y_train)
gb_y_pred = gb.predict(X_test)


# ExtraTreesRegressor
et.fit(X_train, y_train)
et_y_pred = et.predict(X_test)
```

Random Forest Regressor (rf):

The Random Forest Regressor is an ensemble learning method that leverages a collection of decision tree regressors. By aggregating predictions from multiple trees, the Random Forest aims to provide accurate and robust car price predictions.

Gradient Boosting Regressor (gb):

The Gradient Boosting Regressor, unlike Random Forest, builds an ensemble of decision trees sequentially. It starts with a single decision tree and progressively creates additional trees to correct the errors made by the previous ones. This sequential approach often results in improved predictive accuracy.

Extra Trees Regressor (et):

The Extra Trees Regressor is another ensemble learning method, similar to Random Forest. It constructs multiple decision trees to predict car prices. However, what sets it apart is the introduction of additional randomness in the tree-building process, which helps mitigate overfitting.

4.5 MODEL EVALUATION

In this section, we thoroughly evaluate the three regression models: Random Forest Regressor, Gradient Boosting Regressor, and Extra Trees Regressor. Model evaluation is critical in determining which model is most suitable for our specific use case.

Comparison of Model Performance:

Random Forest Regressor:

```
print("r2_score of Random Forest Regressor:", r2_score(y_test, rf_y_pred))
mse = mean_squared_error(y_test, rf_y_pred)
print("Mean Squared Error (MSE) of Random Forest Regressor: ",mse)
mae = mean_absolute_error(y_test, rf_y_pred)
print("Mean Absolute Error (MAE)of Random Forest Regressor: ",mae)
```

```
r2_score of Random Forest Regressor: 0.8341469523824485

Mean Squared Error (MSE) of Random Forest Regressor: 24459753.631692447

Mean Absolute Error (MAE)of Random Forest Regressor: 2877.1613220584836
```

The Random Forest Regressor achieved the highest R-squared (R2) score among the models, indicating that it explains 83.41% of the variance in the target variable. This suggests that it provides the best overall fit to the data.

With a Mean Squared Error (MSE) of 24,459,753.63 and Mean Absolute Error (MAE) of 2,877.16, it demonstrates relatively accurate predictions with low error metrics.

This model is well-suited for applications where predictive accuracy is crucial, and it has shown robust performance.

Gradient Boosting Regressor:

```
print("r2_score of Gradient Boosting Regressor :", r2_score(y_test, gb_y_pred))
mse = mean_squared_error(y_test, gb_y_pred)
print("Mean Squared Error (MSE) of Gradient Boosting Regressor: ",mse)
mae = mean_absolute_error(y_test, gb_y_pred)
print("Mean Absolute Error (MAE) of Gradient Boosting Regressor: ",mae)
```

```
r2_score of Gradient Boosting Regressor : 0.7021549884616032

Mean  Squared  Error  (MSE)  of  Gradient  Boosting  Regressor:  43925726.462724455

Mean Absolute Error (MAE) of Gradient Boosting Regressor: 4542.273816637462
```

The Gradient Boosting Regressor achieved an R2 score of 0.7022, which is slightly lower than the Random Forest model, indicating that it explains 70.22% of the variance in the target variable.

It has a higher MSE (43,925,726.46) and MAE (4,542.27) compared to the Random Forest model, suggesting that predictions may not be as precise.

Extra Trees Regressor:

```
print("r2_score of Extra Trees Regressor:", r2_score(y_test, et_y_pred))
mse = mean_squared_error(y_test, et_y_pred)
print("Mean Squared Error (MSE) of Extra Trees Regressor: ",mse)
mae = mean_absolute_error(y_test, et_y_pred)
print("Mean Absolute Error (MAE)of Extra Trees Regressor: ",mae)
```

```
r2_score of Extra Trees Regressor: 0.8156274338249531

Mean  Squared  Error  (MSE)  of  Extra  Trees  Regressor:  27190983.88522658

Mean Absolute Error (MAE)of Extra Trees Regressor: 2726.9931811838846
```

The Extra Trees Regressor demonstrated an R2 score of 0.8156, suggesting that it explains 81.56% of the variance in the target variable.

It has an MSE of 27,190,983.89 and a relatively low MAE of 2,726.99. This model combines the strengths of both Random Forest and Gradient Boosting.

The Extra Trees model is well-suited for applications where predictive performance and computational efficiency are important.

CHAPTER 5

PROJECT INSIGHTS AND FUTURE PROSPECTS

5.1 SUMMARY

- Our project aimed to predict the prices of used cars based on various features and attributes. We utilized a comprehensive dataset for analysis and modeling.

- During the project, we performed extensive data preprocessing and feature engineering to ensure the quality and relevance of our dataset.

- Three regression models were evaluated: Random Forest Regressor, Gradient Boosting Regressor, and Extra Trees Regressor. These models were selected for their potential to provide accurate predictions in a regression task.

- After comprehensive model evaluation, our primary model recommendation is the Random Forest Regressor, which achieved the highest R-squared score and demonstrated relatively accurate predictions with low error metrics.

5.2 KEY FINDINGS

Data Quality and Preprocessing

The success of our project relied heavily on the quality and preprocessing of the dataset. Clean and well-structured data ensured the reliability of our analysis. We employed data cleaning techniques, such as handling missing values, removing outliers using the Interquartile Range (IQR) method, and filtering out irrelevant records. This step significantly improved the accuracy and stability of our predictive models.

Additionally, label encoding was applied to categorical features to convert them into numerical representations, allowing seamless integration with machine learning models.

Feature Engineering

Feature engineering played a critical role in enhancing model performance. We carefully selected and engineered relevant features, including:

- **Year** – Vehicle manufacturing year, impacting depreciation.
- **Manufacturer & Model** – Brand and model influence market value.
- **Fuel Type** – Fuel efficiency and demand variations affect pricing.
- **Odometer Reading** – Higher mileage generally reduces vehicle value.
- **Title Status** – Indicates ownership and vehicle condition.

  By ensuring relevant features were included and unnecessary ones removed, we enhanced the ability of our models to make precise predictions.

Exploratory Data Analysis (EDA) Insights

- **Fuel distribution analysis** revealed that gasoline was the most common fuel type, followed by diesel.

- **Manufacturer distribution** indicated a strong market presence for certain brands, affecting their resale value.
- **Correlation analysis** highlighted strong relationships between variables, allowing better model optimization.

Model Evaluation

The model evaluation results demonstrated the strengths and weaknesses of each regression model:

- **Random Forest Regressor** outperformed others with an R-squared score of 0.8341, meaning it explains 83.41% of the variance in car prices.
- **Gradient Boosting Regressor** and **Extra Trees Regressor** showed competitive performance but had slightly lower accuracy.
- The Mean Squared Error (MSE) and Mean Absolute Error (MAE) metrics further validated the robustness of our models.

  We also observed that ensemble models, particularly Random Forest and Extra Trees, handled non-linear relationships in the data more effectively than traditional regression models.

Recommendations

- For applications requiring high predictive accuracy, we recommend using Random Forest Regressor, given its robust performance and ability to generalize well across different data distributions.
- Further enhancements can be made by integrating deep learning models or incorporating additional features such as location-based pricing to improve model performance.
- Deploying the model as a web-based tool could provide real-time predictions for users, increasing accessibility and usability.

5.3 LIMITATIONS AND FUTURE WORK

- The challenges and constraints faced during the project are potential biases in the dataset or limitations in feature engineering.

- The current model is trained on a dataset of 8,200 car listings, which may not fully represent global or region-specific markets. Expanding the dataset to include a larger variety of car brands, regions, and additional listings could enhance the model's generalizability and robustness.

- We could explore the use of deep learning models for price prediction or investigate the impact of location-based features on pricing, which will be possible directions for future research or extensions of the current project.

CHAPTER 6

CONCLUSION

In this project, we embarked on a journey to predict the prices of used cars, leveraging the power of data analysis and machine learning. Through meticulous data preprocessing, feature engineering, and the evaluation of multiple regression models, we have gleaned valuable insights and achieved noteworthy results.

By carefully selecting specific attributes, we were able to create predictive models that offered substantial accuracy and precision in estimating car prices. Our primary recommendation, the Random Forest Regressor, excelled with an impressive R-squared (R2) score of 0.8341, meaning it explains 83.41% of the variance in the target variable. Its low Mean Squared Error (MSE) and Mean Absolute Error (MAE) metrics further validate its efficacy.

In closing, our project demonstrates the immense potential that data analysis and machine learning offer in the domain of used car pricing prediction. The road ahead may have its challenges, but with a commitment to quality and innovation, we are poised for further exploration and growth in this exciting field of data Science.