Department of Computer Science

Used Automobile Price Analysis/Prediction

CS714 – Big Data Analytics & Cloud Computing

Guided By:
Dr. Lisa Fan

Prepared By:
Jaydeep Prajapati - 200468915
Vrunda Kakadiya - 200471747

Table of Content

1. Introduction:

In today's modern era, transportation is one of the major factors in the country's economy. Cars are one of the easiest modes of transportation. Businesses that provide services like buying and selling cars to customers are worth around 1.6 trillion USD and are expected to grow at a rate of 6.1 in the upcoming few years. The used car price prediction is beneficial for the organization whose business is driven by the car-selling model. Even though cars are the easiest mode of transportation, not everyone can purchase a new one due to their affordability and people are more willing to buy a used car. Many companies use data analysis to get a better level of understanding about car distribution according to cities where they can accelerate their sales. As the demand for used cars increases every day, the data generated around the world have been exploding in size and complexity, and the massive amount of data produced every day surpasses the capacity of conventional processing systems, requiring us to adopt cutting-edge computing infrastructures capable of handling parallel and distributed processing. It is incredibly difficult to efficiently mine such vast volumes of data, and to do so in a timely manner necessitates the development of platforms that are more complex. Using quick, dependable, and scalable computational architecture, big data infrastructures have emerged to address the issue of big data analytics. They offer excellent quality attributes like elasticity, availability, and resource pooling as well as the capability of on-demand and simple self-services. Because a large number of automobiles are bought and sold, predicting accurately the price of cars is a pain nowadays. As the solutions to issues, machine learning models have been implemented based on these huge records. However, using machine learning techniques on large and complicated datasets is computationally expensive and uses a lot of logical and physical resources, including CPU, memory, and data file space. To address this problem, there are many big data analytics tools and services that are provided by cloud computing platforms such as Amazon AWS, Google Cloud Platform, and Microsoft Azure.

2. Problem Statement

The main goal of this project is to create a machine-learning model for independent businesses to estimate the cost of a customer's automobile right from the history of used car data. With such a large amount of data on used cars available in the market, the challenge is to do an intensive analysis of historical data of customers' buying patterns to accurately forecast automobile prices and to offer the best recommendations to the end users. The records for used cars worldwide are so huge in volume that the traditional approach is not capable enough to analyze the big data, instead, it requires a cloud platform for the storage, analysis of data, and prediction of some good results based on the records with the use of Machine Learning models.

3. Examples of the problem

Even though there are around 15,000 dealership companies in Germany for cars, a person can't find a perfect car. Hence, the main aim of this project is to predict the price of a car based on its features by analyzing the data which provides a better understanding of used cars to distribution companies. This ambition is beneficial for a person as well to find a car as to their requirement in a specific budget. For example, the same cars with the same features might be available at different prices on the platform which might cause a loss to either customers or companies as the price has been decided based on a few features only.

4. General Solution

The project mainly focuses on the analysis of big data on the cloud computing platform, as well as recommendations using analysis results and prediction of the car prices for the end users. Microsoft Azure provides many services for the smooth analysis of big data such as Azure Synapse Analysis, Databricks, HDInsight, and Machine Learning Studio. The project will be constructed using Apache Spark on the HDInsight service. Different machine learning regressor models such as Linear Regression, Decision tree, Random Forest, Gradient Boosting, and XGBoost can be used to predict the price of cars precisely out of which some models will be implemented in the project. Meaningful data analyzed in various algorithms and outcomes will be predicted with the RMSE (Root Mean Square Error) and R2_score that reflects the accuracy of the models. As the final stage of the solution, a dashboard on Microsoft Power BI will be created to visualize different analyses of the data in order to generate recommendations for the users.

5. Approach

General Approach

In this project below key points will be explained and discussed further in deep,

- Explanation of dataset
- Microsoft Azure and reasons for selecting it
- Architecture of Solution
- Storage
- HDInsight and Apache Spark
- Machine Learning Approach
- Visualization

5.1. Dataset Explanation(Background)

This data has been scraped from different websites for a year which contain valuable information about cars. The dataset of classified ads for cars contains approximately 3.5 records of the used car for sale in Germany and Czech Republic since 2015.

Feature	Description
Maker	Maker of a car
Model	Model of car
Mileage	Mileage of a car
Manufacture_year	Year of manufacture of a car
Engine_displacement	Displacement of the engine in ccm
Engine_power	Power of engine in kW
Body_type	Type of a car
Color_slug	Color of a car
Stk_year	Year of last emission control
Transmission	Automatic or manual
Door_count	Number of doors
Seat_count	Number of seats

Fuel_type	Type of fuel
Date_created	Date of scrapped record
Date_last_seen	Date of the last ad
Price_eur	Price of the car in EUR

5.2. Microsoft Azure platform and reasons for selecting it

Cloud computing has been an essential component in the big data analytics world. Cloud services providers provide different services IaaS(Infrastructure As a Service), PaaS(Platform As a Service), and SaaS(Software As a Service) for the storage of big data and computations. There are a number of mature cloud providers such as Amazon AWS, Microsoft Azure, IBM, and Google that can deliver a wide array of big data services. [5] The Microsoft Azure cloud places a strong emphasis on analytics and AI services. Processing large amounts of structured and unstructured data are simple with the Azure platform. A fully managed infrastructure that comprises Azure database services, analytics services, machine learning, and data engineering solutions is also included, as well as real-time analytics. Azure offers a huge selection of analytics-related products and services. The most well-liked services at the moment are HDInsight, Databricks, Synapse Analysis, and Azure Analysis Services. So, these are some of the many reasons Microsoft Azure HDInsight was used to build the solution.

5.3. Architecture of Solution

The below diagram shows the architecture of the solution. It mainly involves Storage for efficiently storing the big data, HDInsight cluster, and Big Data tools for processing large data and building machine learning models to predict the car price using the pre-processed data. The prediction result and data will be again stored back in the storage account. Finally, it includes the Power BI tool for the visualization of different analyses of the data. All these stages will be explained in brief further.

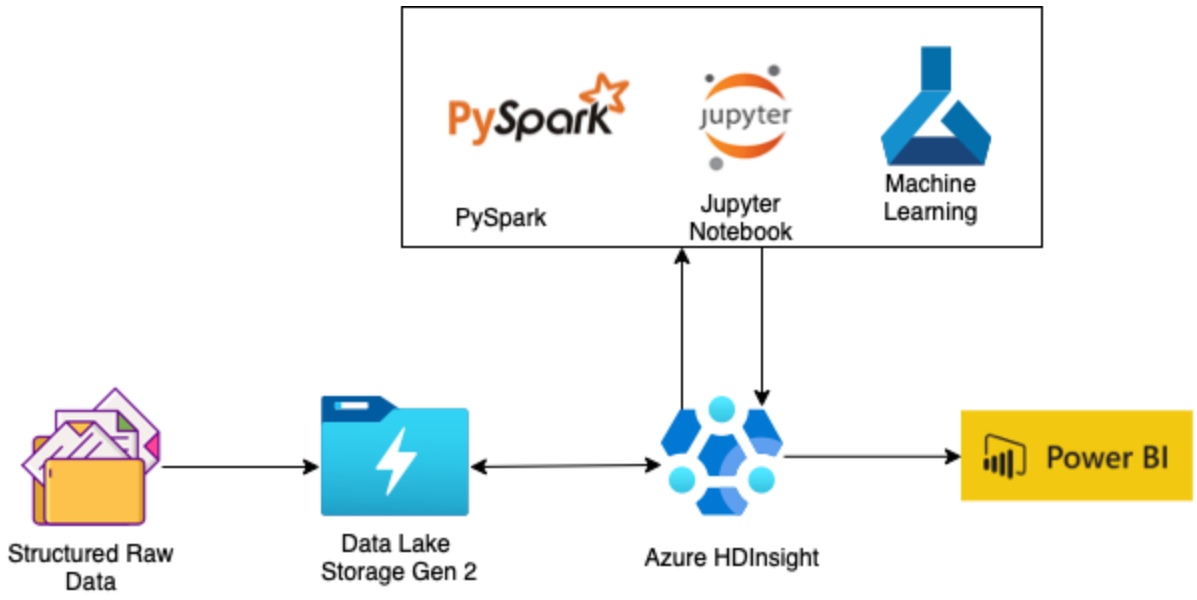

Fig. 1 Architectural Diagram

5.4. Storage (Data Lake Storage Gen 2)

Azure Storage provides highly accessible, massively scalable, durable, and secure cloud storage for a wide range of data items. The storage includes Azure Blob storage and Data Lake Storage Gen 1 and Data Lake Storage Gen 2. Data Lake Storage Gen1 should not be used for any new projects as it became a legacy service. Azure blob storage is designed to hold large volumes of unstructured data. It has a flat namespace structure This feature has a substantial influence on performance, particularly in big data analytics settings. The inclusion of a hierarchical namespace to Blob storage is a critical component of Data Lake Storage Gen2. It is the combination of both Azure Blob Storage and Data Lake Storage Gen1. It was created with big data analytics in mind, and it is an essential component of current data analytics, data science, and data warehousing infrastructures.

Azure Data Lake Storage Gen 2 accounts are created by selecting the "Enable hierarchical namespace" option when creating an Azure Storage Account. After creating ADLS Gen 2 upload the CSV file to the container of the storage account to process it using big data tools.

5.5. HDInsight and Apache Spark

Azure HDInsight is one of the most popular analytics services of Microsoft Azure. HDInsight is an enterprise analytics solution that is interoperable with prominent platforms like Spark, Kafka, and Apache Hadoop which connects to the services like SQL Data Warehouse and Azure Data Lake that make it simple to implement analytical

pipelines. HDInsight can interface with custom analytic tools and supports a wide range of popular programming languages, including Python, JavaScript, R,.NET, and Scala. Spark provides in-memory cluster computing primitives. A Spark job could load, cache, and query data in memory frequently. In-memory computing is far quicker than disk-based programs such as Hadoop, which exchange data via the Hadoop distributed file system (HDFS). Apache Spark in the HDInsight cluster basically works on 3 nodes(Virtual Machine) Head Node, Zookeeper Node, and Worker Node, and the number of nodes and size can be configured as per the requirement of the project and the cost varies as per the selection of virtual machines.

The below diagram shows how spark runs in the HDInsight cluster.

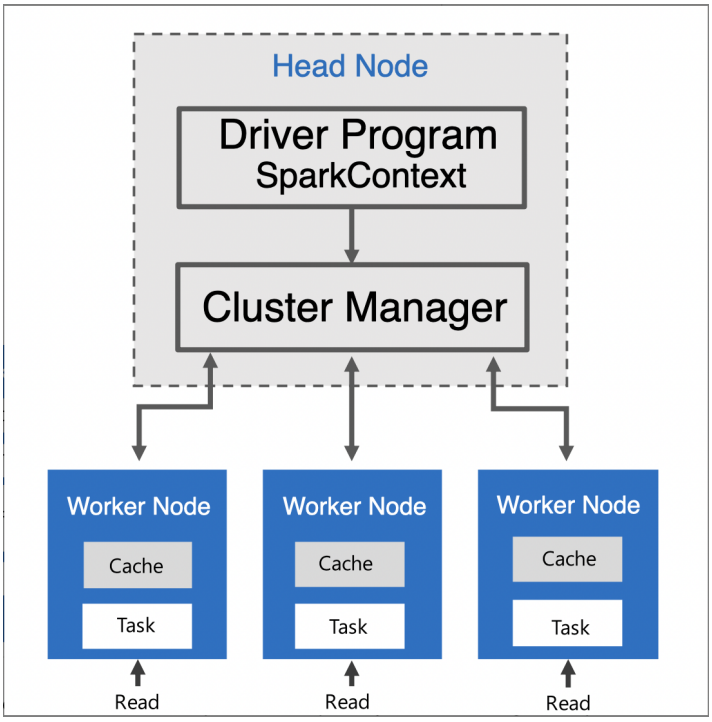

Fig. 2. Apache Spark HDInsight Cluster [2]

The SparkContext executes the user's main function and all parallel computations on the worker nodes. The SparkContext then collects the outcomes of the operations. The Hadoop distributed file system is read and written to by the worker nodes.

The resource for HDInsight spark v2.4 was created in Azure using a Data Lake Storage Gen 2 storage account. In the project, 2 head nodes, 3 ZooKeeper nodes, and 3 worker nodes were used. It took around 20-30 minutes to create a cluster with all the necessary tools and libraries pre-installed. Jupyter notebook was created in order to Extract,

Transform and Load the used cars sample data using PySpark and to perform analysis and price prediction.

5.6. Machine Learning Approach

- **Pre-processing**
  - In order to make data more useful and accurate, null and irrelevant records needed to be either removed or substituted with other values.
  - These are the steps taken to clean the data
    - Dropped null values from columns - ‘Maker’, ‘Model’
    - Mileage needed to be fixed between 500 to 500000.
    - Dropped records where Manufacture Year is less than 1980
    - Dropped records with prices less than 3000 euros and greater than 200000 euros
    - Forward, Backward Fill, and Interpolation for the Fuel Type, Transmission, Door Count, and Seat Count.
    - Typecasting from string to integer

- **Feature Scaling**
  - There are 16 columns available in the dataset but not all are necessary for the machine-learning model.
  - So, by using correlation table dropped some columns and selected these features for price prediction system - 'maker', 'model', 'mileage', 'manufacture_year', 'transmission', 'door_count', 'seat_count', 'fuel_type', 'price_eur'
  - Non-numerical features were converted into numeric data types by Label Encoding as models only accept the numeric values.
  - The Data set is divided into the standard ratio of 80% and 20% for training and testing purposes respectively.

- **Multiple Linear Regression**
  - Multilinear regression finds the relationship between all independent variables to the dependent variable.
  - $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \epsilon$
    - Where for i=n observations:
    - y_i = dependent variable
    - x_i = explanatory variables
    - β_0 = y-intercept
    - β_p = slope coefficients
    - ϵ = Error term (residuals)

- **Random Forest**
  - A Random forest can be used for classification and regression. In this scenario, it's been implemented as a regressor as price prediction is a continuous value.
  - This algorithm works based on the ensemble technique, where the same algorithm runs multiple times to avoid any differences in results.
- **Gradient Boosting**
  - Gradient boosting relies its best on the next step, by merging previous trees to models which reduces the chances of an error.
  - This algorithm always adds one tree at a time and adds residuals to it to make the model more accurate by boosting the gradient.

5.7 Visualization

Data visualization tools and techniques are critical in the Big Data era for better analyzing hugh volumes of data and making data-driven decisions since data is increasingly used for significant management decisions. Tableau, Power BI, and Google Charts are some popular visualization tools for big data.

After the pre-processing of data, cleaned data was stored in the form of a table in the blob container of the storage account. These are some analyses from different graphs created using the pre-processed data from storage.

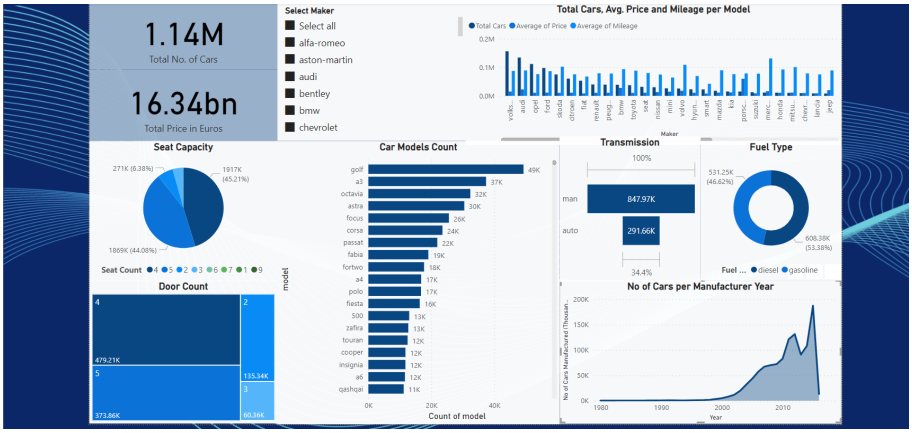

Fig. 3. Power BI Dashboard

Fig. 3 represents a Power BI dashboard that contains different graphs and plots that represent the analysis of different data fields such as the number of cars, makers, models, price, etc… It gives some recommendations based on these fields. Some of the recommendations and analysis can be predicted below.

- **Count of cars based on their door capacity**
  The highest numbers of cars are four doors.

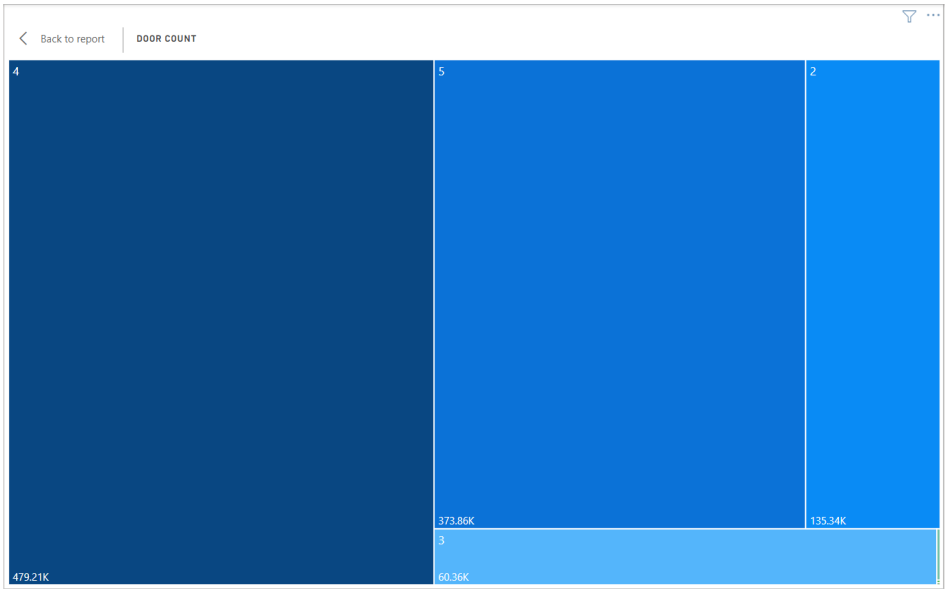

Fig. 4. Count of Vehicles based on Door Count

- **Count of vehicles based on the fuel type**
  Vehicles with Gasoline fuel type dominated the count of vehicles

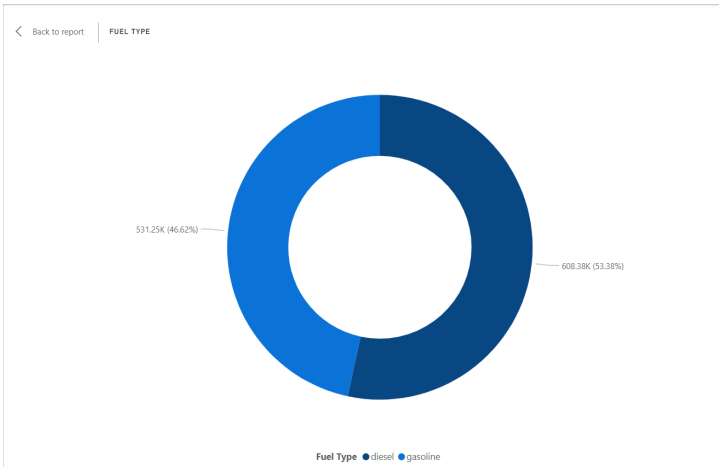

Fig. 5. Count of Vehicle-based on fuel type

- **Count of cars per manufacture year**
  Most of the records of the car are manufactured from 2005 to 2015 with 2013 year getting the highest numbers of products for the cars.

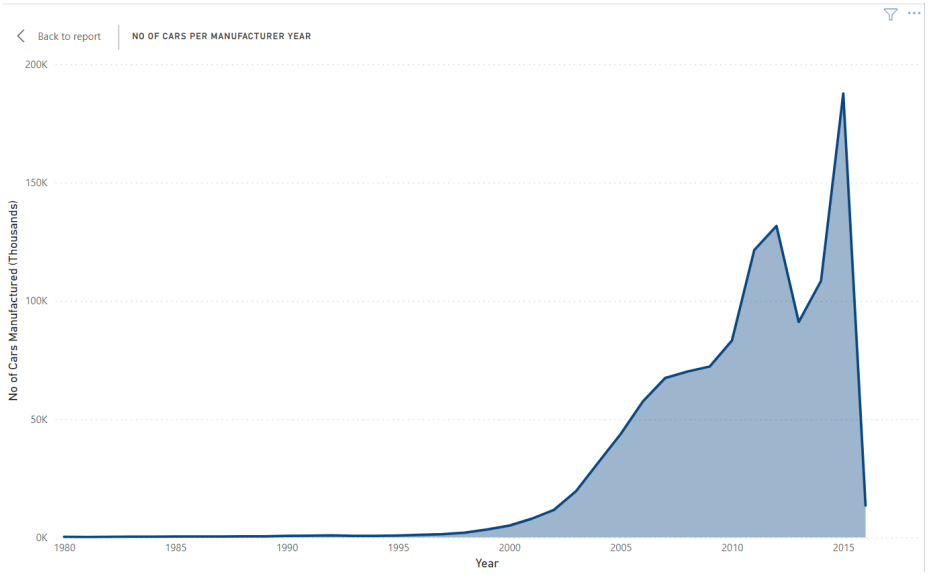

Fig. 6. Count of Vehicle per year of manufacture

- **Count of cars based on model**
  The top 5 popular automobile models are - golf, a3, Octavia, Astra and focus

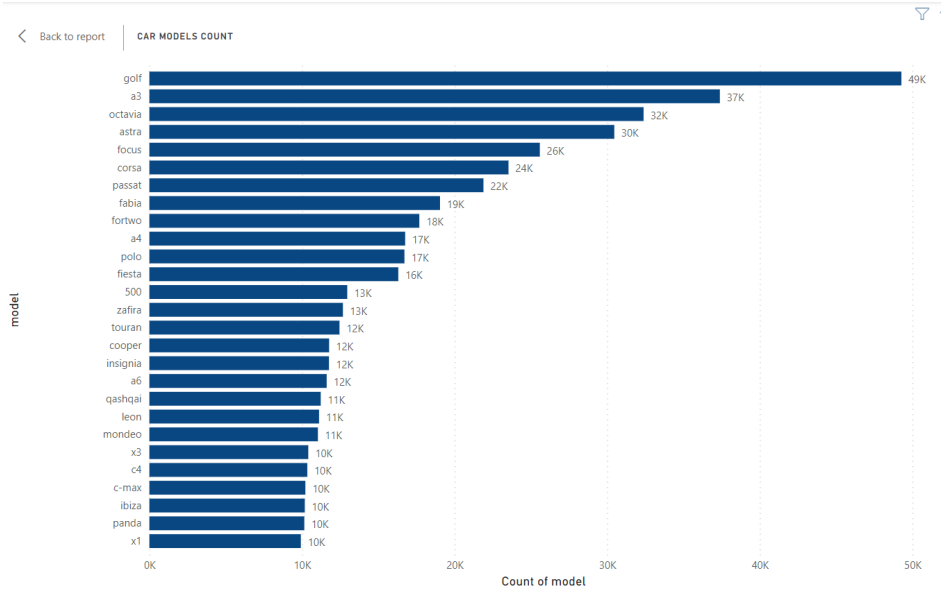

Fig. 7. Count of Vehicle-based on models

- **Count of vehicles, avg price, and mileage per model**
  The top five car manufacturers based on the count of vehicles are - Volkswagen, Audi, Opal, Ford, and Skoda. After filtering based on avg price, the top 5 models are - Lamborghini, Tesla, Bently, Aston-Martin, and Porche.

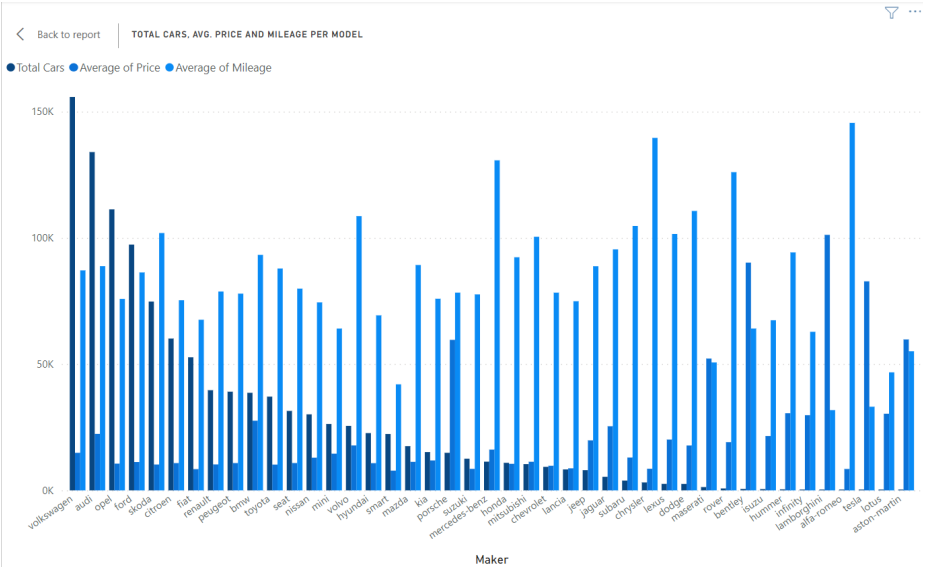

Fig. 8. Count of vehicles, avg price, and mileage per model

- **Count of Cars based on Seating Capacity**
  Over 80% of the cars have a seating capacity as of four or five

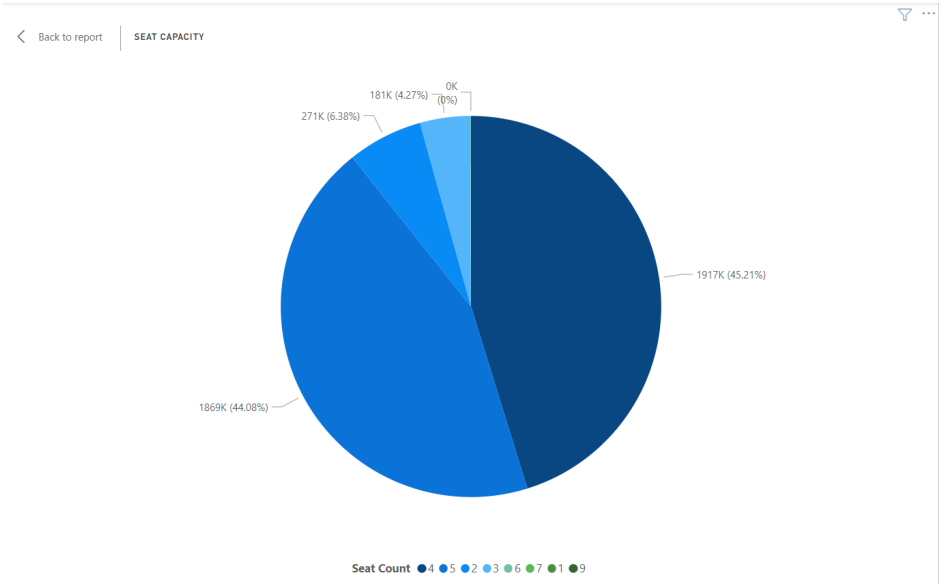

Fig. 9. Count of Vehicles based on Seating Capacity

- **Count of Vehicles based on transmission**
  Analysis shows manual cars dominate the market with 75% of total records.

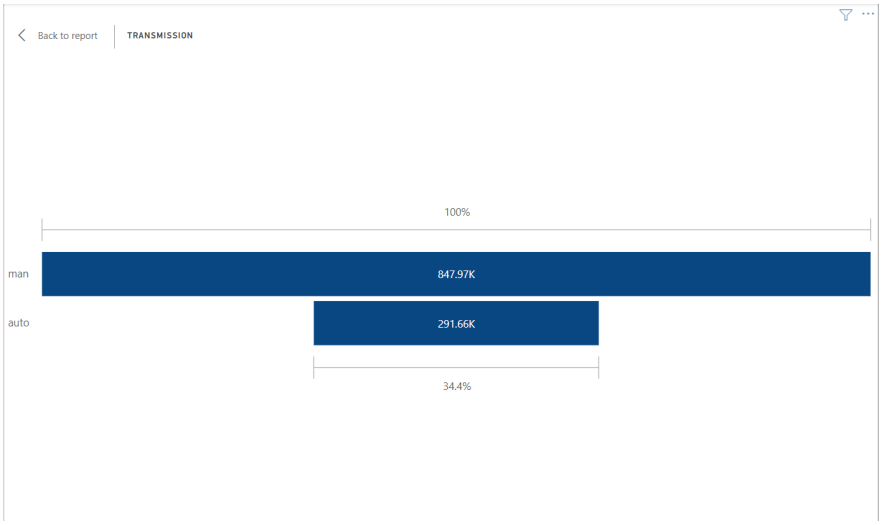

Fig. 10. Count of vehicles based on Transmission

6. Results

All previously explained models perform different processes and predict different values for the same input. The results for all three implemented machine-learning models are

	RMSE	R2 Score
Linear Regression	10525.08	0.33
Random Forest	4391.79	0.88
Gradient Boosting	4049.56	0.86

In comparison to the above three models, Linear regression is not performing well in prediction. It has only 33% of accuracy. On the other hand, Random Forest is able to achieve the highest of 88% accuracy with the 4391.79 RMSE. Gradient Boosting has the lowest RMSE of 4049.56 with 86% accuracy. Random Forest is the best choice when accuracy is the main concern as it is too time-consuming compared to other models and companies should choose Gradient Boosting when accuracy is as important as time.

7. Discussion of Relevant Literature

1. **Usage of Hadoop and Microsoft cloud in big data analytics: An exploratory study [3]**
   This case study investigates how Hadoop and Microsoft cloud are employed in big data analytics and how the big data sector is changing. It also examines these activities, as well as how programs address large amounts of data analytics and resolve concerns about growing data volumes. It also discusses big data tools that are currently supported by Microsoft Azure such as Azure HDInsight, Databricks for Azure, Azure Steam Analysis, HDFS, and MapReduce.

2. **Price Prediction for Pre-Owned Cars Using Ensemble Machine Learning Techniques [4]**
   This study was for predicting fair prices for pre-owned cars in the Mumbai, India region using various machine learning models. In the paper, two different machine learning algorithms Decision Tree and XGBoost were implemented, and before implementation data pre-processing and feature selection were performed. The study concluded that both models perform very well and give high accuracy scores.

8. Limitations and Possible Extensions

The major limitation of this project is that the model is implemented only on the data of two countries, which is limited to those cities or countries' set of rules and requirements. Hence, the same machine learning model can not be applicable to another city of the same or different countries.

Beneficial for dealers if the same thing is implemented on an automated pipeline for the ETL process. Microsoft Azure's Synapse analytics might be helpful to perform in HDInsights. Building a deep learning model would be beneficial mainly in the cases where dealers already knew the specific requirements such as black, white, and blue colors are the most favorite colors for cars, and prices for those models are slightly high in comparison to others. Hence, building a deep learning model with some predefined weights is beneficial to the dealer in terms of pricing.

9. Conclusion

Creation, configuration, updating, and deletion of the HDInsights spark cluster are easy to build with the Azure portal. Jupyter notebook is pre-installed with azure which accelerates the implementation rather than focusing on the configuration between python files and azure. In terms of accuracy and model selection, random forest and gradient boosting give better accuracy however, the time complexity of the random forest is way higher compared to gradient boosting. Analysis of specific data in PowerBI is quite helpful for the dealer to recommend the cars to customers.

10. References

[1] Classified Ads for Cars | Kaggle
[2] What is Apache Spark - Azure HDInsight | Microsoft Docs
[3] USAGE OF HADOOP AND MICROSOFT CLOUD IN BIG DATA ANALYTICS: AN EXPLORATORY STUDY
[4] Price Prediction for Pre-Owned Cars Using Ensemble Machine Learning Techniques
[5] Azure Big Data: 3 Steps to Building Your Solution