Jagannath University Research Journal (JURJ) Volume No.-II, Issue No.–II, November, 2021, ISSN: 2582-6263

# Study of E-bay Used Cars Sales using Tableau

**Nainika Kaushik, Dr. Manjot Kaur Bhatia**
[a]Jagan Institute of Management Studies, Rohini Sector-5,New Delhi-110085,India, [b]Jagan Institute of Management Studies, Rohini Sector-5,New Delhi-110085,India

**Abstract: Since Germany has such a large used automobile industry, numerous adverts for sales are posted on eBay. We used eBay data to provide information about the vehicle, including its make, brand, type, horsepower, and kilometers. The data consists of almost 370000 rows and approximately 20 columns. We are also given the price at which the car was sold, in addition to these characteristics. Our goal is to optimize sales by understanding the impact of the factors and the price. As a result, our primary task is to determine the significant factors that influence pricing. In order to work on sales, we also work to: identify the elements that affect the price of a used car; identify the factors that impact the price of a used car; and identify the factors that influence the price of a used car. Finally, understand and analyze the relationship between the elements that determine the price of the car, and utilize the information to forecast the price of the car when all of the factors are known.**

**Keywords: Tableau, Classification, Decision Tree, Random Forest, Support Vector Machine,K-means**

## I. INTRODUCTION

Given that Germany has vast sales of used cars, many advertisements are posted on eBay. We have taken this data from eBay, where there are factors related to the car like its make, brand, type, power, kilometers used, etc. Along with these factors, we are also given the price at which the car was sold. Understanding the impact of the factors and the price, our objective is to maximize the sales. Hence, our main aim is to find the significant factors influencing the price. Understanding the impact of the factors and the price, our objective is to maximize the sales.

This dataset aims to determine the depreciation of used cars over the years and predict the value of the used car based on the data. The database consists of Over 370000 used cars scraped from eBay-Germany and has 21 columns.

Having a colossal dataset is a responsibility as well as an opportunity. There is a lot to derive from it, but at the same time, it needs more time and effort on the preprocessing. With more than 370,000 rows, we modified and normalized, and retained 340,000 rows. These rows were further used for modeling and analysis.

Using the data provided, our objective is to:

- Identify the factors which influence the price of a used car.
- Understand and analyze the correlation between the factors that influence its price.
- Use the information to predict the price of the cars given all the factors.

Our results are focused on three main factors:

- Accuracy –We check the model accuracy using the R square value.
- Error – Every model gives an amount of error along with accuracy. It is essential to take that into account as well. We check the RMSE value to check the error.
- Complexity – While we may achieve higher accuracy with each modeling, the complexity increases exponentially. Maintaining a decent accuracy, we also need to ensure that the complexity of the model we choose is a good trade-off.

## II. METHODS AND APPROACH

To address our objective and reach a conclusion, we use Tableau and Python. We will use this tool for data preprocessing, exploration, and modeling.

We are following the SEMMA approach, for each part, we are subdividing into:

**S – SAMPLE:**Instead of using the original data

directly, we use the subset per our requirement. It helps retain the relevant information, which increases the chances of better predictions.

**E – EXPLORE:** To decide our predictors and target, it is essential that we first understand the relationship of all the factors among themselves and then decide what should be chosen.

**M- MODIFY:**Before we use our predictors for modeling, we must modify them as per the needs. At times, the data is highly skewed or has inconsistent data, or may even have missing values; it requires modification so that the chosen predictors can drive the best results.

**M- MODEL:**Once we are done with sampling, exploring, and modifying the data, here comes the part of modeling. According to the type of predictors and target, we choose the possible models that can fit our data.

**A – ASSESS:**After creating models, we compare and use the one that best suits our business requirements and gives the most relevant result.

### *2.1 DataDescription*

The original data set has 371,538 rows, and 19 columns carry the data about used cars available for sale in Germany. This dataset describes the listing activity and the metrics of car sales on eBay Germany. The data is crawled from the advertisements posted on the eBay website that included all the information from dates to brands. Out of 19 variables, we will choose one as our target variable-Price while others as our potential predictor variables.

This data file contains all of the information required to learn more about insights, geographical availability, and the metrics required to make predictions and draw conclusions. This dataset is part of Kaggle, and the source can be found on Kaggle.com.

- dateCrawled:whenthisadwasfirstcrawled,all field-valuesaretakenfromthisdate
- name:"name"of thecar
- seller:private ordealer
- offerType
- price:thepriceonthe adtosell thecar
- abtest
- vehicleType
- yearOfRegistration:atwhichyearthe carwasfirstregistered
- gearbox
- powerPS:powerof thecarinPS
- model
- kilometer:howmanykilometersthe carhasdriven
- monthOfRegistration:atwhichmonth the carwasfirstregistered
- fuelType
- brand
- notRepairedDamage:if thecarhasadamagewhichisnotrepaire dyet
- dateCreated:the dateforwhich thead atebaywas created
- nrOfPictures : number of pictures in the ad (unfortunately this field contains everywhere a0and isthususeless(bugincrawler!))
- postalCode
- lastSeenOnline:when thecrawlersawthisadlastonline

### *2.2 Exploring Data*

To understand our data and their relationship, we need to use different visualization aids like histograms and bar charts.
We have created various individual and several variables analysis charts on tableau: (i) Price v/s Vehicle Type v/s Power PS
In the graph below (Fig.1), it is clear that SUVs have the highest mean price and highest power. Power PS v/s vehicle type v/s Mean Price It is clear that cars having power PS in the range of 171-700 have the highest mean price for each vehicle type. Also, for coupe, limousine, small cars, and station wagon, the mean price is almost equal for the power PS between the range of 132 to 171.

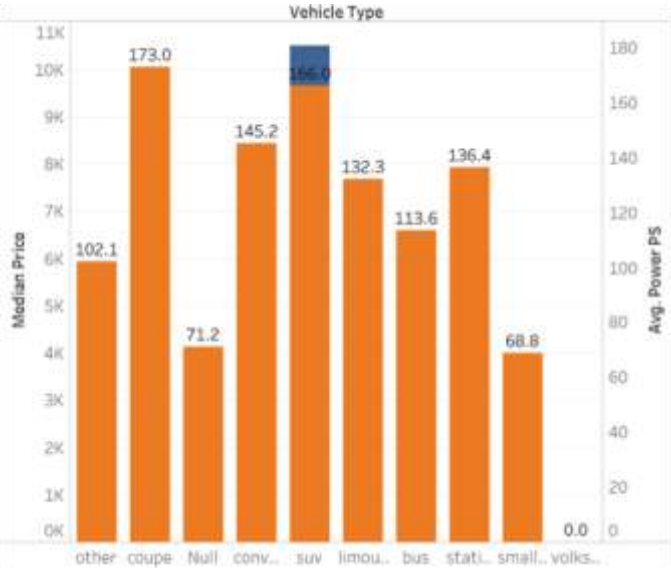



**Fig 1. Pricev/sVehicleTypev/sPowerPS**

(ii). Brand v/s Fuel Type v/s Price

While comparing which brand with different fuel types has the highest price, it is shown that Porsche with hybrid fuel has the highest price, followed by Volvo with hybrid fuel. (Fig.2)

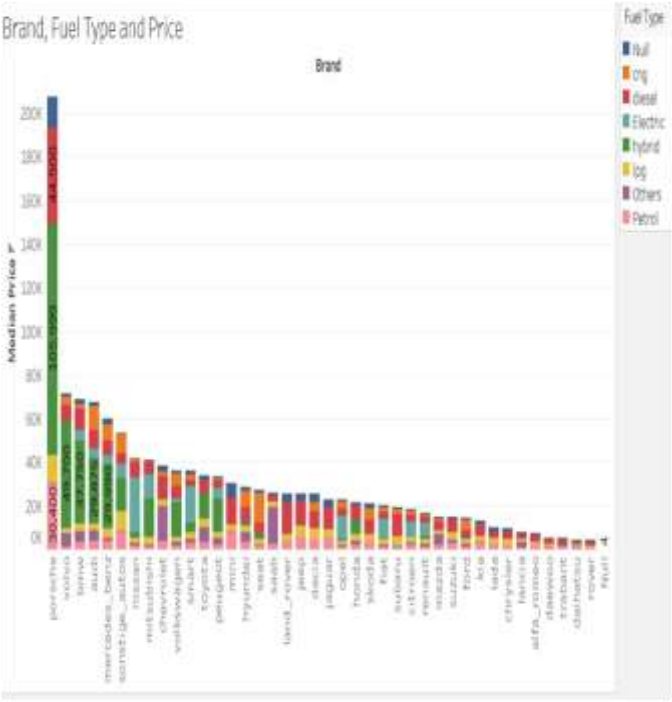



**Fig 2. Brand v/s Fuel Type v/s Price**

(iii) Fuel type v/s gearbox v/s Power

LPG cars with automatic gear have the highest power, followed by Diesel cars. (Fig.3)

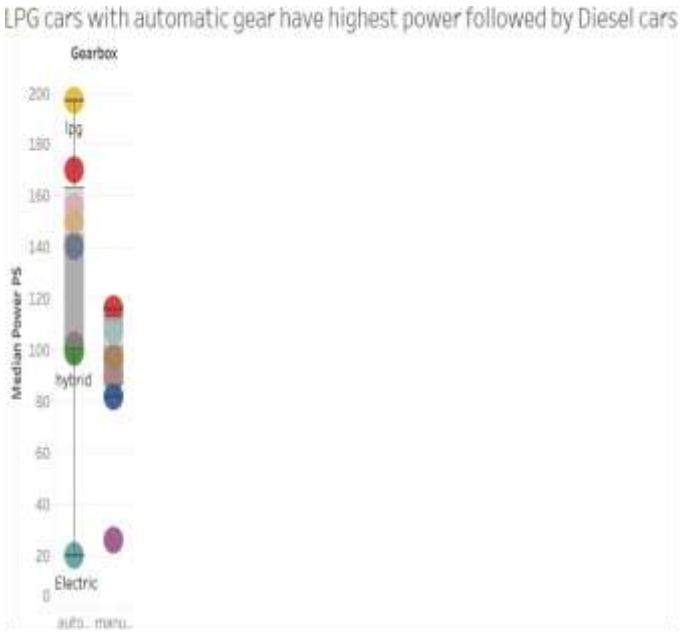



**Fig 3. Fuel type v/s gearbox v/s Power**

### *2.3 Data Cleaning and preprocessing*

Data cleaning is the process of detecting and correcting corrupt or inaccurate records from the data set. It refers to identifying incomplete, incorrect, inaccurate, or irrelevant parts of the data and then replacing, modifying, or deleting them from the data set. It includes Removing Null Values, Removing Duplicates, and Missing Data Patterns. In the missing data pattern, we have chosen nine predictors. We will look for the rows where the missing values are more significant than three and exclude them, as they won't be beneficial for predicting the target.

## III. MODELLING

We used Machine Learning techniques to predict used car prices from other dependent variables.

We considered three types of errors for evaluating the model: Mean Absolute Error, Mean Squared Error, and RMSE (Root Mean Squared Error).

### *3.1 SupervisedLearningAlgorithms*

1. Linear Regression - Linear regression is a common

technique in statistical data analysis. It is used to determine how a dependent variable and one or more independent variables have a linear relationship.

```
from sklearn import metrics
print(metrics.mean_absolute_error(y_test,y_pred))
print(metrics.mean_squared_error(y_test,y_pred))
print(np.sqrt(metrics.mean_absolute_error(y_test,y_pred)))

```

```
3755.751358122828
92084890.36109316
61.284185220355404
```

2. Logistic regression – It is a statistical model that, in its basic form, uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression estimates the parameters of a logistic model (a form of binary regression).

```
from sklearn import metrics
print(metrics.mean_absolute_error(y_test,y_pred))
print(metrics.mean_squared_error(y_test,y_pred))
print(np.sqrt(metrics.mean_absolute_error(y_test,y_pred)))

```

```
5517.32525
499182070.0585
74.27869984053302
```

The regression section shows that linear regression fits better than logistic regression in our dataset.

3. Decision Tree - The decision tree constructs regression or classification models in a tree structure. It incrementally divides a dataset into smaller and smaller subsets while also developing an associated decision tree. The result is a tree with leaf nodes and decision nodes.

```
from sklearn import metrics
print(metrics.mean_absolute_error(y_test,y_pred))
print(metrics.mean_squared_error(y_test,y_pred))
print(np.sqrt(metrics.mean_absolute_error(y_test,y_pred)))
```

```
4620.051366983528
187111345.89939764
67.97095973269414
```

Decision trees can work with both categorical and numerical data.

4. Random forest - Random forests, also known as random decision forests, are an ensemble learning method for classification, regression, and other tasks that works by training a large number of decision trees and then outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.[1] Random decision forests compensate for decision trees' proclivity to overfit their training set.

```
from sklearn import metrics
print(metrics.mean_absolute_error(y_test,y_pred))
print(metrics.mean_squared_error(y_test,y_pred))
print(np.sqrt(metrics.mean_absolute_error(y_test,y_pred)))

```

```
2831.7797593575688
47529468.58979034
53.214469454816225
```

5. K Nearest Neighbors Regression - The K nearest neighbors algorithm is a simple algorithm that stores all available cases and predicts the numerical target using a similarity measure (e.g., distance functions)[2]. The average of the numerical targets of the K nearest neighbors is a simple implementation of KNN regression. Another method is to take an inverse distance weighted average of the K closest neighbors. The exact distance functions are used in KNN regression as in KNN classification.

```
from sklearn import metrics
print(metrics.mean_absolute_error(y_test,y_pred))
print(metrics.mean_squared_error(y_test,y_pred))
print(np.sqrt(metrics.mean_absolute_error(y_test,y_pred)))
```

```
3942.6261458333333
77168579.0934375
62.79033481224108
```

6. Support-vector machines – Supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis [3]. An SVM training algorithm constructs a model that assigns new examples to one of two categories, resulting in a non-probabilistic binary linear classifier given a set of training examples, each of which has been labeled as belonging to one of two categories (However, there are methods for using SVM in a probabilistic classification setting, such as Platt scaling). An SVM model in the form of the examples as points in space that have been mapped in a form that the examples from various categories are separated by as large a gap as possible. [4] New examples then are mapped into that same space and classified based on which side of the gap they fall on. The metrics: 4943.067916666667,107167538.3558334 and 70.30695496653703.

7. XGBoost (eXtreme Gradient Boosting) - Is a popular machine learning algorithm. It is useful for supervised learning tasks like regression, classification, and ranking. [5] It is based on the gradient boosting framework's principles and is intended to "push the extreme of machine computation limits to provide a scalable, portable, and accurate library."

```
from sklearn import metrics
print(metrics.mean_absolute_error(y_test,y_pred))
print(metrics.mean_squared_error(y_test,y_pred))
print(np.sqrt(metrics.mean_absolute_error(y_test,y_pred)))

```

```
3348.45425
45190081.62475
57.86582972705049
```

Supervised Learning gives fewer errors, which means the supervised algorithm fits well on our dataset. Random Forest offers the best result from all the classification algorithms applied. Gradient Boosting and KNN also work well.

### *3.2 UnsupervisedLearningAlgorithms*

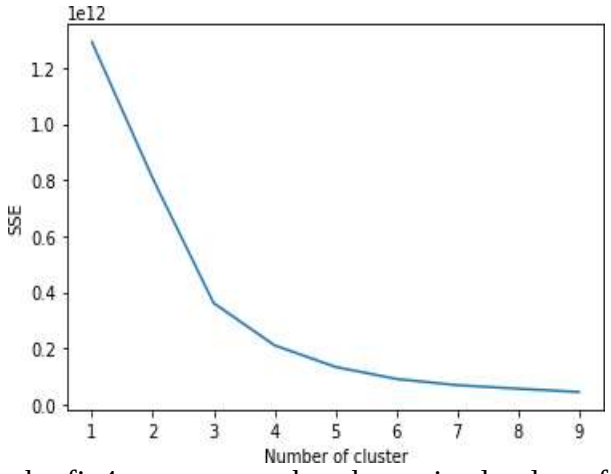
1. Kmeans: K-means clustering is one of the simplest and most popular unsupervised machine learning algorithms. A cluster is a collection of data points that have been aggregated together due to certain similarities.[6] You need to define a target number k, which refers to the number of centroids you need in the dataset. A centroid is an imaginary or actual location representing the cluster's center.

2. Elbow method K-means – Is a simple unsupervised machine learning algorithm that groups data into a specified number (k) of clusters. The elbow method performs k-means clustering on the dataset for a range of k values (say, 1-10) and then computes an average score for all clusters for each value of k.[7] The Elbow method is used to get the optimal number of clusters by fitting the model with a range of values for K.

```
from sklearn import metrics
print(metrics.mean_absolute_error(y_test,y_pred))
print(metrics.mean_squared_error(y_test,y_pred))
print(np.sqrt(metrics.mean_absolute_error(y_test,y_pred)))
```

```
5834.09875
120671495.93729167
76.3812722465396
```

**Fig 4. K-means**




From the fig4 we can see that the optimal value of the cluster is 4. So we have applied the K-Means algorithm for a number of clusters, i.e., 4. And it is

visible after looking at errors that an unsupervised algorithm is the worst fit for our dataset.

## IV. CONCLUSION

Out of all the models validated, Random Forest has the lowest error and is best considered for analyzing this dataset.

Predicting modeling aims to help the business grow by providing insights into what is going on now and how it can be improved. Here are some insights that can be worked up to enhance the business growth rate.

1. The graphs and data of the top 5 states in respect to their count as well as average selling price (Fig 5.):

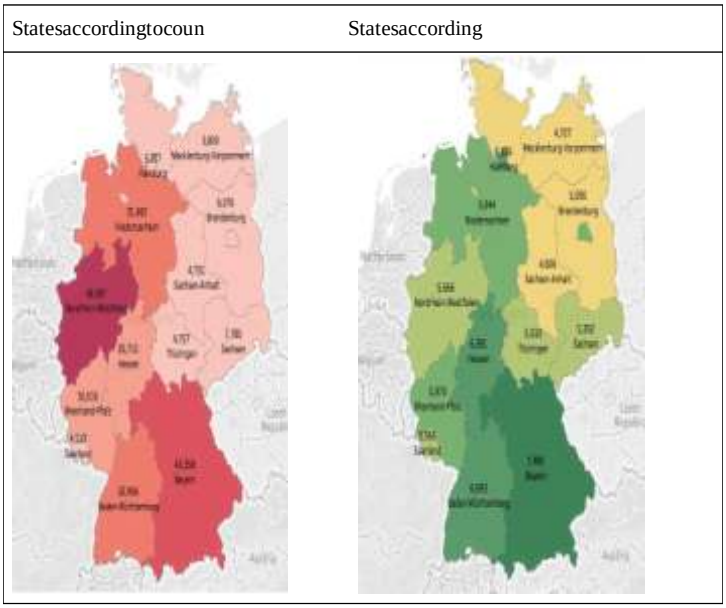



**Fig 5. States in respect to their count as well as average selling price**

(a) As seen from the data above, the three states occurring in the top 5 tables of states count are also in the top 5 states of highest average price. These states are Bayern, Baden-Württemberg, and Hessen. Therefore, to maintain their positions, eBay should keep their marketing continue.

(b) While Berlin is not among the top 5 states in car count, it comes at the 5th at an average price. Given that it is also the capital of Germany, there is a vast scope of improvement if the work is done in the right direction, and its count can also improve.

2. We also focused on the top 5 brand and vehicle type combinations with the highest prices. (Table 1.)

**Table 1- Top 5 combinations of brands and vehicle type with highest prices.**

| Brand | Vehicle Type | Year of Registration | Max. Price |
|---|---|---|---|
| Audi | Coupe | 2016 | 150,000 |
| Mercedes Benz | Convertible | 1969 | 148,000 |
| Skoda | Station wagon | 2012 | 145,000 |
| Audi | Convertible | 2015 | 138,980 |
| Mercedes Benz | Convertible | 1953 | 130,000 |

The top 5 years in respect to maximum price are - 2016, 1969, 2012, 2015, and 1953. The reason for 2012-2016 is that being the latest models of Audi and Skoda, their reselling price is high. So while the occurrence of 1969 and 1953 might seem to be an outlier, it is not.

Given that Germany has a vast following and interest in the historic car, the price is relatively high.

## REFERENCES

[1]. Arthi J., Akoramurthy B. (2018) Extrapolation and Visualization of NPA Using Feature Based Random Forest Algorithm in Indian Banks. In: Pattnaik P., Rautaray S., Das H., Nayak J. (eds) Progress in Computing, Analytics and Networking. Advances in Intelligent Systems and Computing, vol 710. Springer, Singapore. https://doi.org/10.1007/978-981-10-7871-2_58

[2]. Kaliszyk C., Urban J. (2015) FEMaLeCoP: Fairly Efficient Machine Learning Connection Prover. In: Davis M., Fehnker A., McIver A., Voronkov A. (eds) Logic for Programming, Artificial Intelligence, and Reasoning. LPAR 2015. Lecture Notes in Computer Science, vol 9450. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-662-48899-7_7

[3]. Kaushik N,Bhatia M.K.,Rastogi S.(2020)SVM and cross validation using RStudio. International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249-8958, Volume-10 Issue-1

[4]. Kaushik N,Bhatia M.K.,Rastogi S.(2020)Various data classification technique on crime dataset. International Journal of Management, Technology And Engineering, Page No : 84-92 DOI:16.10089.IJMTE.2020.V10I01.20.35460

[5]. Kaushik N., Bhatia M.K. (2020) Information Retrieval from Search Engine Using Particle Swarm Optimization. In: Sharma H., Govindan K., Poonia R., Kumar S., El-Medany W. (eds)

Advances in Computing and Intelligent Systems. Algorithms for Intelligent Systems. Springer, Singapore.https://doi.org/10.1007/978-981-15-0222-4_11

[6]. Kaushik N., Bhatia M.K. (2022) Twitter Sentiment Analysis Using K-means and Hierarchical Clustering on COVID Pandemic. In: Khanna A., Gupta D., Bhattacharyya S., Hassanien A.E., Anand S., Jaiswal A. (eds) International Conference on Innovative Computing and Communications. Advances in Intelligent Systems and Computing, vol 1387. Springer, Singapore. https://doi.org/10.1007/978-981-16-2594-7_61

[7]. Guo, J. (2018). Developing a Visualization Tool for Unsupervised Machine Learning Techniques on* Omics Data (Doctoral dissertation).

## AUTHOR'S BIOGRAPHIES

**Nainika Kaushik** She is working as an assistant professor in department of information technology, JIMS. She has completed her M.Tech. from IGDTUW with the specialization in mobile and pervasive computing and B. Tech. from GGSIPU in computers science and engineering. She has published many papers in international journals. She has attended and presented papers in various national and international conferences. She has received the Award titled "Proud Indian - Innovative Scientist Award 2022 in Computing." by International Multidisciplinary Research Foundation (IMRF) , NGO Registered with NITI Aayog NGO Darpan & Govt of A.P, HQ: Vijayawada India

**Dr. Manjot Kaur Bhatia:** She is working as a Professor in the Department of Master of Computer Application, Jagan Institute of Management studies, Sector-5, Rohini, Delhi. She has done M.C.A., M.Phil. and Ph.D. from University of Delhi in the field of Information Security. She has more than 17 years of teaching experience. She is actively involved in teaching and research in the areas of Security, Databases, Linux, Operating System and Information Security. She has been guiding the students in their project work and motivating them towards the research work also. Her other areas of interest includes Cloud Computing, Steganography, Data Hiding, Information security and Software testing. She has published many research papers in various refereed International journals and has presented number of papers in international conferences. She has been the session chair in various international conferences.