Used Car Analysis

1 Study Purpose

A dataset with information on used vehicles is provided. Various characteristics on car sales are available such as car "model", "manufacturer", "year", "condition" of the vehicle etc.

The ultimate objective of this work is to determine which of these features significantly contribute to the definition of a car price based on the knowledge resulting from the provided dataset. Different Artificial Intelligent techniques will be employed allowing to study the considered problem from different aspects.

Hereafter, the organization of this document is presented.

A brief introduction to the principal steps handling the considered problem is presented in § 2. In § 3 the data preparation is discussed necessary phase for any further data employment. A discussion of the three considered approaches and the related outcomes regarding the feature importance is developed in § 4.

In § 5 the principal points of the presented work are resumed.

Finally, § 6 develops some points of interest regarding future work extension.

2 Problem Approach

Aiming at defining the crucial characteristics of a given dataset determining the price of a used car, a first systematic study is proposed comprised of the following main steps:

- data preparation
- model development for data exploitation
- acquirement of the desired response through analysis of the resulting model outcomes.

There are multiple artificial intelligence techniques to handle this problem. Each possible alternative is associated with advantages and inconveniences. For this reason, the present work attempts to access this issue from different perspectives. Each one will be briefly explained before presenting the resulting outcomes responding to the considered problem.

However, no definite conclusion should be drawn as the major purpose of this work is to introduce a methodology able to satisfy the required business objective. Due to time constraints and computer limitations a short number of experiments are examined. The provided conclusions appraise only the exploited trials. A more exhaustive investigation of the developed methodology, involving a wider range of numerical values, is necessary for a deeper appraisal of the related feature classification resulting from each suggested model (see also § 6).

3 Data Treatment

A dataset comprised of $426,880$ entries or samples is provided. Each sample contains information for eighteen different characteristics or features. A short description of each available feature is presented hereafter.

1. id: the sample identifier
2. region: the related geographical region (e.g. Auburn etc.)
3. price: the vehicle price
4. year: the entry year
5. manufacturer: the manufacturer of the vehicle (e.g. Chevrolet)
6. model: the model of the vehicle (e.g. f-150 xlt)
7. condition: the condition of the vehicle (e.g. good)
8. cylinders: the number of cylinders
9. fuel: the fuel type (e.g. gas)
10. odometer: the miles traveled by the vehicle
11. title_status: the title status of vehicle (e.g. clean)
12. transmission: the transmission of vehicle (e.g. automatic, other)
13. VIN: the vehicle identification number
14. drive: the type of drive (e.g. 4wd)
15. size: the size of the vehicle (e.g. full-size)
16. type: the generic type of vehicle (e.g. pickup)
17. paint_color: the color of the vehicle (e.g. silver)
18. state: the state of listing (e.g. ca).

However, among the available samples, there are entries for which many features are not associated with a value. Various techniques can be utilized for handling such datasets. In this study two approaches are adopted for dealing with missing values.

- The first one keeps $61,005$ entries among the $426,880$ provided ones and sixteen out of eighteen features which are presented as follows:

    1. region
    2. price
    3. year
    4. manufacturer
    5. model
    6. condition
    7. cylinders
    8. fuel
    9. odometer
    10. title_status
    11. transmission
    12. VIN
    13. drive
    14. type
    15. paint_color
    16. state.

- The second approach handling missing values maintains $389,604$ samples but only ten features which are:

    1. region
    2. price
    3. year
    4. manufacturer
    5. model
    6. fuel
    7. odometer
    8. title_status
    9. transmission
    10. state.

Remark 1

- *Feature "id" was advisedly ignored as it is a unique value for each entry contributing only to distinguish samples. It does not add any knowledge when defining a vehicle price.*

- *Due to time constraints, the introduced study will consider only the first dataset involving* $61,005$ *entries and sixteen characteristics. However, it is important to examine the second dataset too as it involves a greater number of samples and fewer features.*

4 Accessing Feature Importance

As referred above, there are multiple ways to establish the contribution of each feature to the definition of a vehicle price and thus determine the most important vehicle characteristics.

The contribution level of each employed feature will be determined according to three principal approaches.

A set of five distinct and *adaptable* artificial intelligence models is developed.

Each of the three approaches exploits these five models in a distinct way and selects the best model according to specific criteria. Various criteria can be used for deciding the most adequate approach to a business need such as an error measurement, size of the data etc.

All the proposed schemes are employing *feature engineering* techniques aiming at creating new pertinent features out of the existing ones. After examination, each model selects the feature combination leading to the most reliable results.

Remark 2

An adaptable artificial intelligence model is a model which involves different possible values for some of its characteristics. The best possible values for these characteristics are determined through automated procedures.

4.1 First Approach

This approach directly applies the five developed models in the considered dataset. Since models are *adaptable* the different possible model configurations are examined. The feature importance, that is the contribution of each involved feature in the definition of a used vehicle price, is determined as the outcome of the *best* selected model (among the experimented ones).

For this approach, the *best* model is comprised of 135 features involving the initial fifteen ones as well new characteristics formed from various combinations of the existing fifteen features.

Figure 1 depicts the importance of each feature through the height of the associated bar.

Table 1 shows the feature importance in a decreasing order. The ten most important features are presented starting with the one having the greater contribution to the vehicle price (located in the first row of the table) to the less important (positioned in the last row of the table).

Hence, for this approach feature (year)^2 is the first most important one. Feature year is the second more important feature. Feature year × transmission is the third most important one and so forth.

Table 1: Ten Most Important Features-Approach 1

Variable (Feature)
$(\text{year})^2$
year
year × transmission
transmission
fuel
year × fuel
year × paint_color
paint_color
$(\text{drive})^2$
year × condition

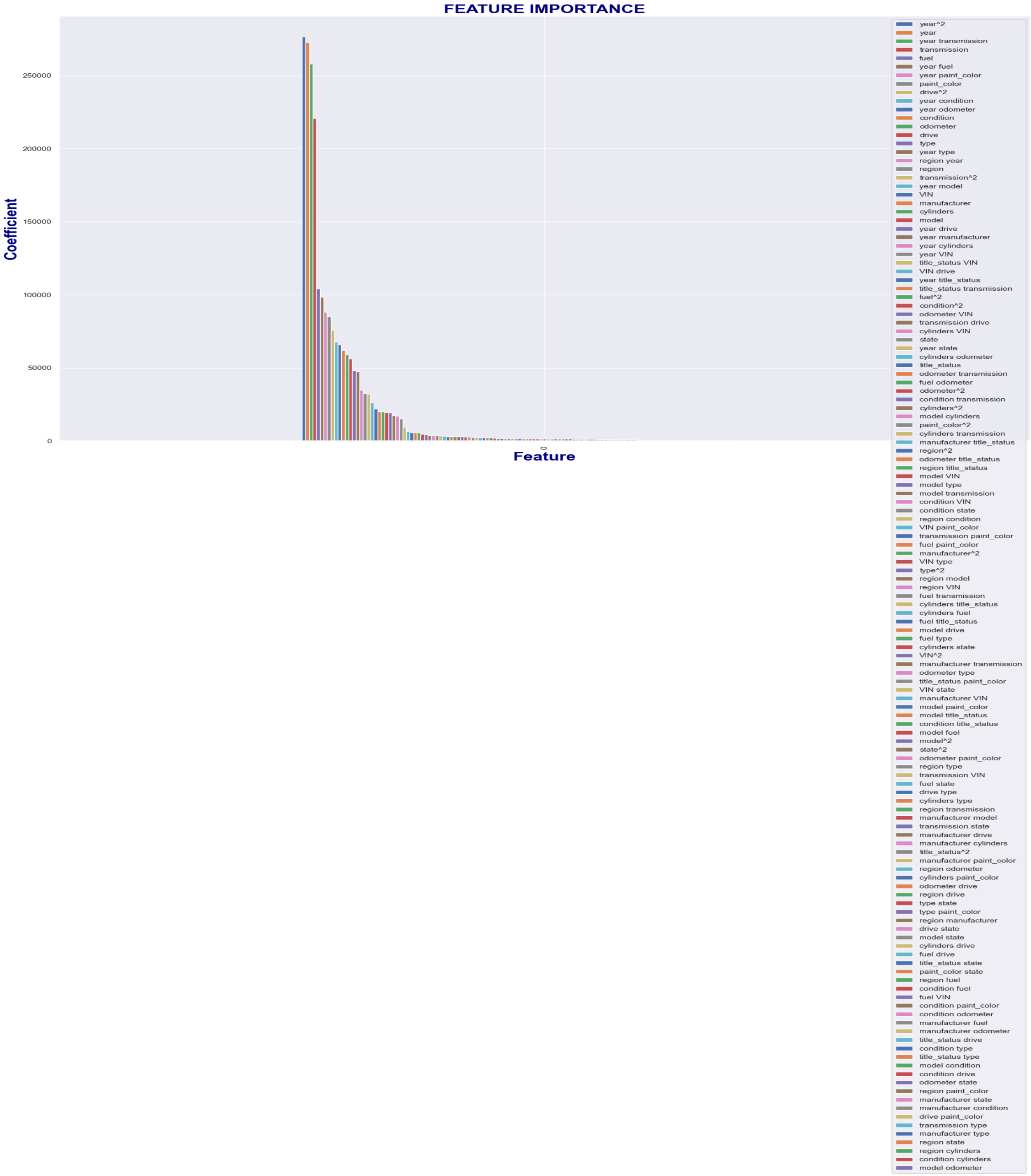

Figure 1: Feature Importance-Approach 1.

Remark 3

Observe how new meaningful features are created by the selected model of the first approach such as (year)^2, year × fuel and so forth. The provided dataset only involves distinct features such as year, fuel but not any combinations of them (see also § 3).

4.2 Second Approach

In this approach, two different techniques are associated in order to define which features highly contribute to the definition of a vehicle price. At a first step, the available data is grouped into three different sets according to specific similarities. Next, for each such group, the features notably contributing to the car price are determined.

Among the available $61,005$ data, the 80% that is $48,803$ entries are used for defining the feature importance. The remaining $12,202$ samples, that is 20% of the total data, are left aside in order to appraise the model performance on unseen information and thus obtain a more objective idea regarding the model evaluation.

4.2.1 Feature Importance in the First Group

The examination of the feature importance for samples classified in the first group, comprised of $13,485$ entries, is going to be studied.

The *best* selected model for this approach engineered 157 features out of the 15 existing ones.

The contribution level of each characteristic is represented through a colored bar in Figure 2. The higher the bar length is the more important the feature is regarding the definition of a vehicle price.

Table 2 illustrates the ten most important features in decreasing order. Thus, for the first group, the most important characteristic is the squared value of "VIN" (the vehicle identification number). The second most most important characteristic is the cubed value of "VIN". The third most important feature is the product of two other features, the "transmission", "VIN" and so on.

Table 2: Ten Most Important Features-Group 1-Approach 2

Variable (Feature)
(VIN)^2
(VIN)^3
transmission × VIN
(condition)^2 × transmission
VIN
odometer × VIN
(odometer)^2
cylinders × odometer
(odometer)^2 × VIN
year × cylinders

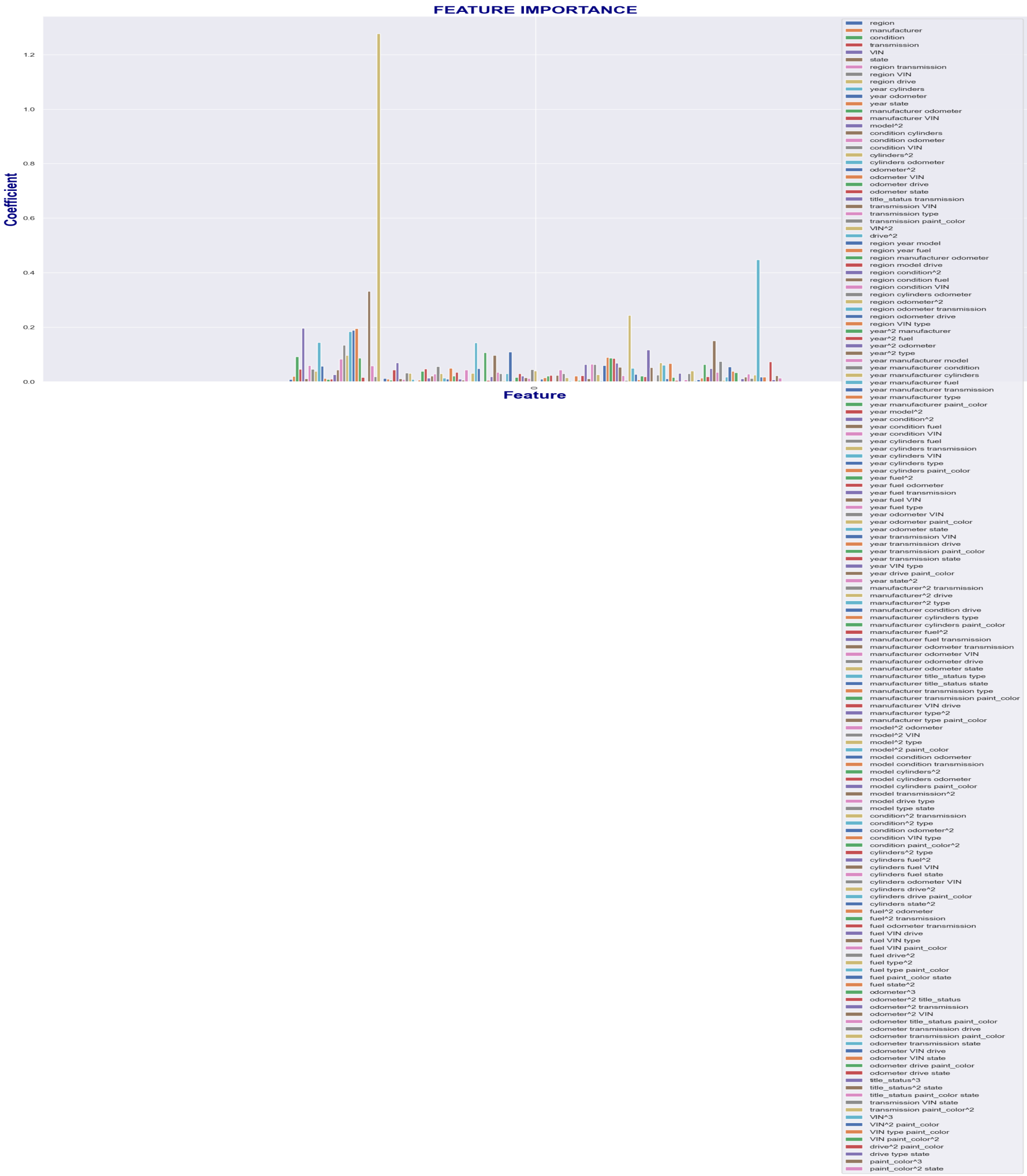

Figure 2: Feature Importance Group 1-Approach 2 (kmeans++).

4.2.2 Feature Importance in the Second Group

The $18,042$ samples belonging in the second group are going to be explored.

The *best* selected model for the second group of samples considered a total of 103 features involving combinations of the existing characteristics.

Figure 3 illustrates the importance of each of the involved features. As previously the height of a bar corresponds to the related importance.

Table 3 depicts the ten most important features in a decreasing order. Hence for this second group of samples, "year" is the most important feature followed by characteristics "VIN" cubed (VIN)^3 and "year" × "VIN" in the second and third order respectively.

Table 3: Ten Most Important Features-Group 2-Approach 2

Variable (Feature)
year
(VIN)^3
odometer
year × VIN
(year)^2
cylinders × (drive)^2
odometer × VIN
(year)^2 × VIN
(model)^3
(model)^2 × VIN

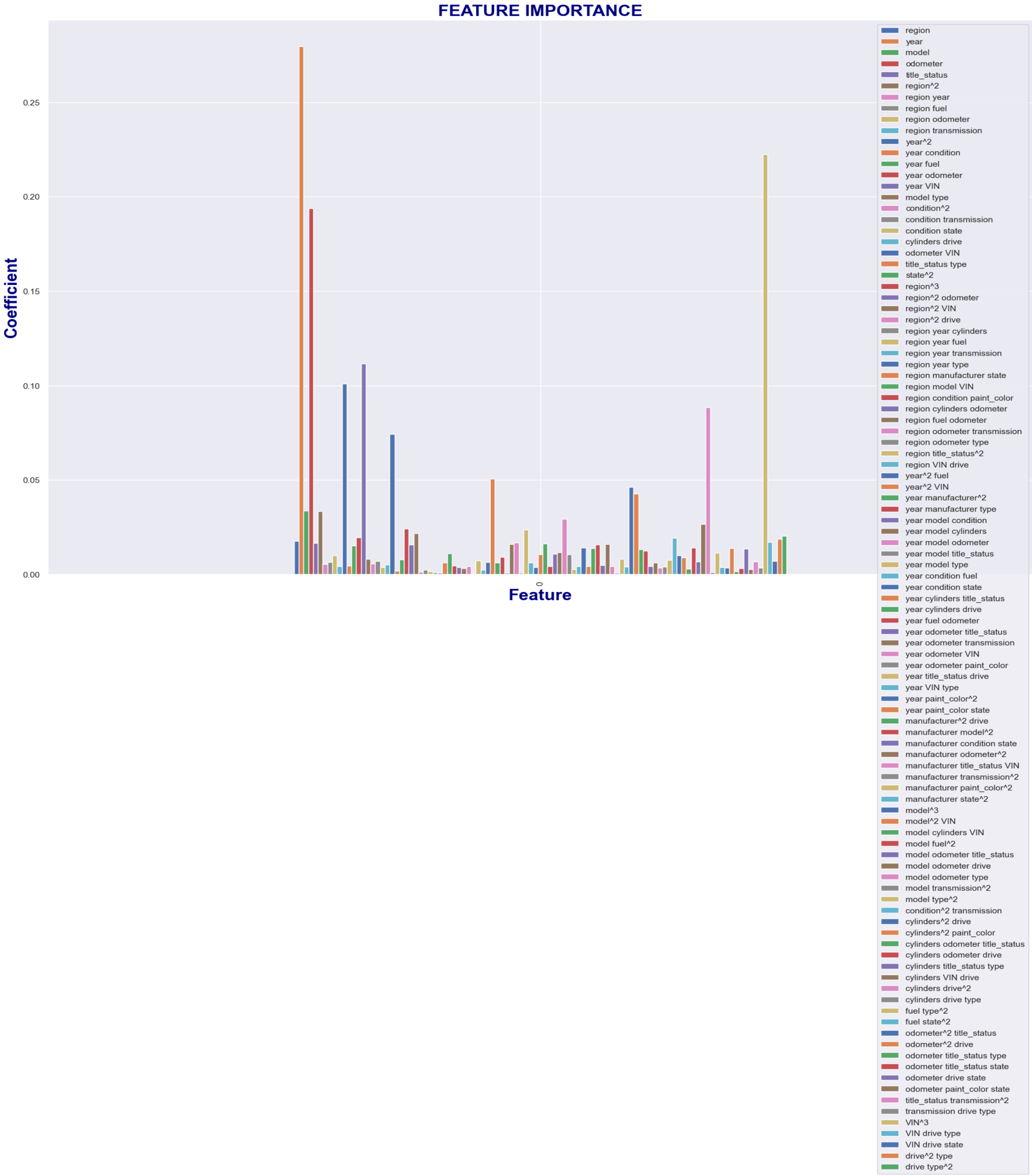

Figure 3: Feature Importance Group 2-Approach 2(kmeans++).

4.2.3 Feature Importance in Third Group

The remaining $17,276$ entries are classified in the third group.

Figure 4 represents the feature relevance through colored bars (a distinct color per feature). The height of each bar effects the related influence of the considered feature.

Table 4 presents the ten most relevant features in a decreasing order.

Feature "year" is in the first place representing the feature influencing at most the price of a vehicle for samples within the third group followed by features "VIN" and "odometer".

Table 4: Ten Most Important Features-Group 3-Approach 2

Variable (Feature)
year
VIN
odometer
$(\text{year})^2$
$(\text{odometer})^2$
year × odometer
model
$(\text{transmission})^2$
$(\text{drive})^2$
transmission

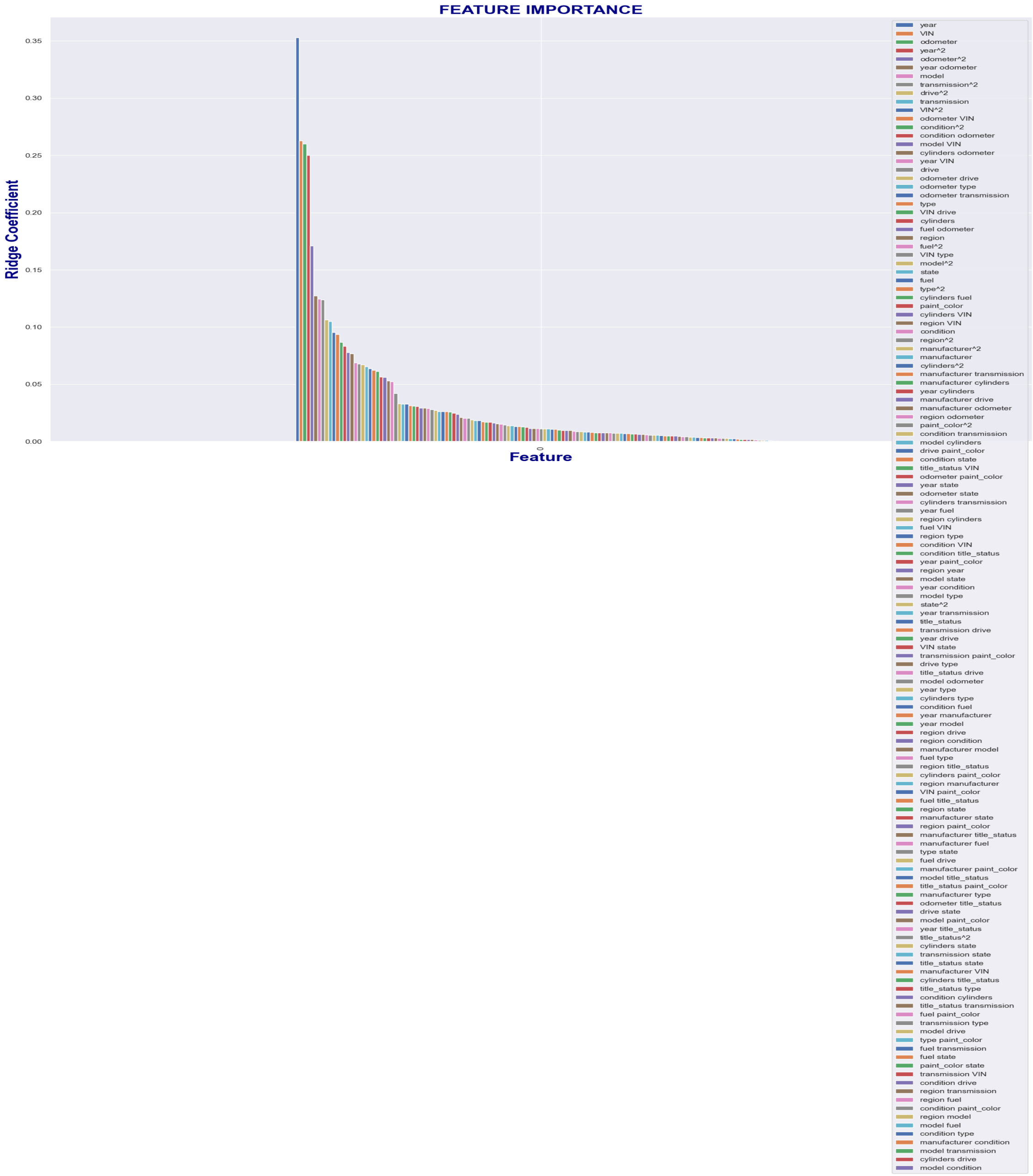

Figure 4: Feature Importance Group 3 (kmeans++).

4.3 Third Approach

With the aim of enhancing the knowledge resulting from the provided dataset, this approach employs a reduced number of new created features named principal components. Each principal component, is a combination of the initial features.

An advantage of this technique consists to an increase of the dataset performance. When dealing with artificial intelligence models, the required number of samples is closely related to the number of features. However, there is no clear meaning for each such feature since different characteristics are involved in its definition (at various rates).

Thus, the fifteen features of the original data are now replaced by ten new features (named PC1, . . . , PC10 where PC stands for principal Component). The values of each sample are now expressed with respect to PC components after adequate computations.

This new dataset is provided to each one of the five models.

The *best* model selected by this approach employs a total 50 PC features created from combinations of the initial ten PC features.

It worths to recall that the *best* model of the previous approaches were employing 135, 157, 103 and 135 features out of the initial fifteen ones.

Figure 5 illustrates the importance of all the 50 PC features through colored bars each one corresponding to a particular feature. The height of each bar indicates the feature contribution to the price of the car.

Table 5 presents in a decreasing order the ten most important features for this approach employing principal components combinations. Hence component PC1 is the most significant one regarding the price definition of a used car followed by PC2 × PC7 as the second most important feature, PC1 × PC7 as the most important third feature etc.

Remark 4

With a detailed analysis, the proportion of the initial features included in the definition of each principal component can be provided. However, this information is not available in this document.

Table 5: Ten Most Important Features-Approach 3

Variable (Feature)
PC1
PC2 × PC7
PC1 × PC7
PC6 × PC7
PC7 × PC8
PC2
$(\text{PC2})^2$
PC2 × PC6
PC7
PC2 × PC4

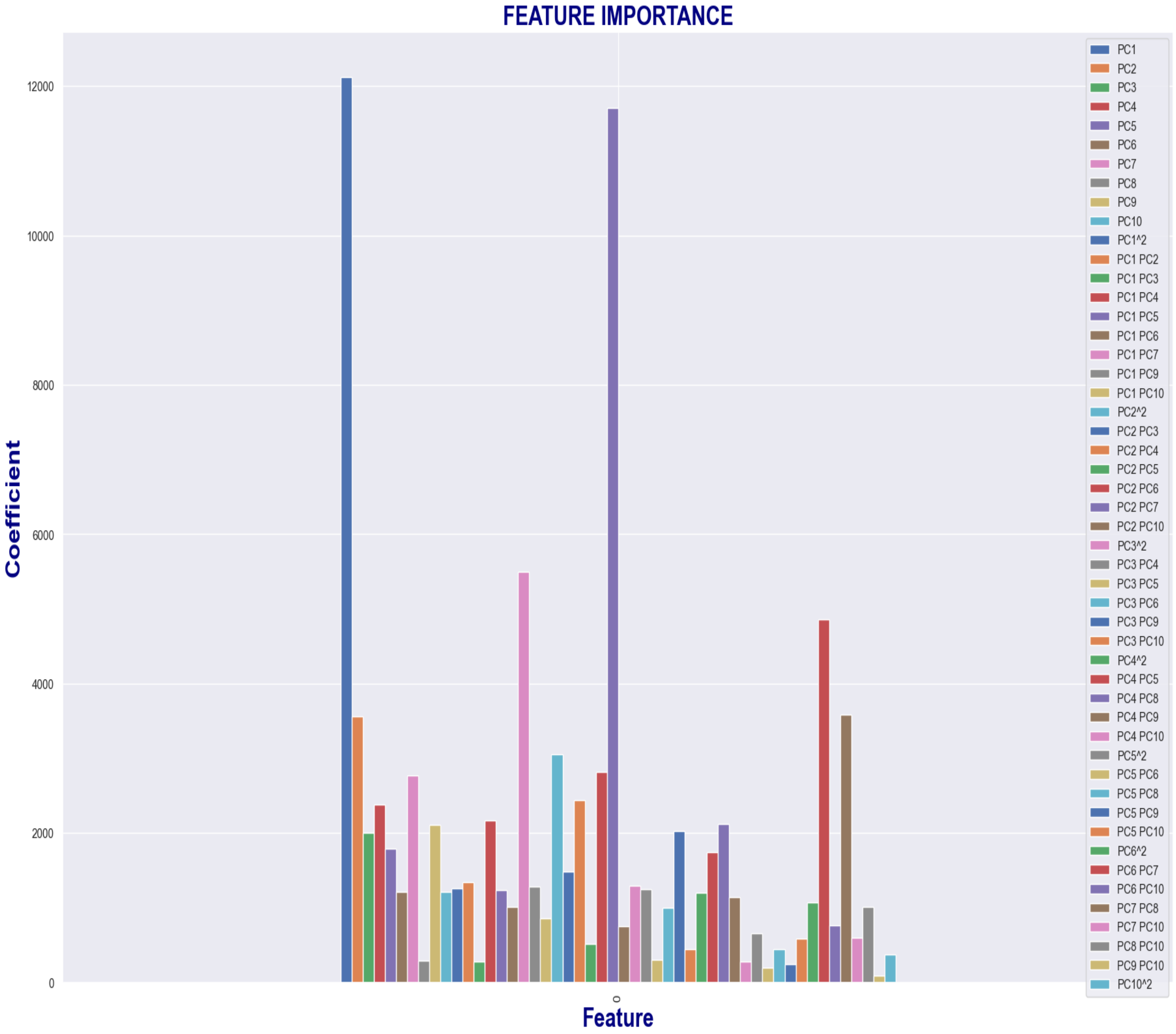

Figure 5: Feature Importance-Approach 3 (PCA).

5 Discussion

The present works introduces a methodology ordering features of a given dataset according to their contribution towards the prediction of the target variable.

More precisely, in this context a dataset comprised of used vehicle characteristics is available. Aiming at predicting the vehicle price, the most significant features should be determined, that is the ones contributing “the most” at predicting the vehicle price.

A first preliminary data treatment maintains sixteen out of the eighteen features (including the target variable) and $61,005$ entries or samples out of the $426,880$ given ones. A second alternative, regarding the data preparation before data exploitation, keeps $389,604$ samples each one providing values only for ten distinct features out of the eighteen initial features.

The current study only refers to the first dataset involving $61,005$ entries and sixteen features.

Feature importance is determined according to three different approaches.

The first approach defines the feature contribution to the price prediction by a direct use of the processed dataset. Different Artificial Intelligence (AI) models are employed each one tested for different hyperparameter values. The best model is selected through an automated procedure. This model provides a feature ordering regarding the associated contribution of each feature to the price definition. Depending upon the model, new features can be created as useful combinations of the fifteen available (the target feature “price” is excluded as this is the one to be predicted). For instance, while feature “year” exists in the provided data, a new feature (year)^2 can be created not initially available.

The second approach is consisted in two parts. Within the first step all data is grouped into subsets regarding particular characteristics determined by AI techniques. Next, for each defined group the most important features are selected (regarding their contribution to the car price prediction). As before potentially new features may be created out of the given ones?

The third approach diminishes the number of features by creating entirely new features as a multiple combination of the initial features. One of the reasons determining the importance of this method is that requires a reduced number of entries. This is an important element especially for the case of the considered dataset which involves only of a small number of the initial entries (due to treatment regarding the missing values). However, the resulting features do not have a direct meaning as they are composed of many other features. Once the new features are determined, just like in the previous two approaches different AI models are tried in association with varying hyperparameter values. The importance of each one of the new features is defined through the best model (defined by the automated procedure).

It is important to precise that the provided feature importance for each of the three approaches does not imply definite conclusions. The presented feature orders are the most *accurate* only regarding the experimented models and hyperparameter values.

A more exhaustive study involving more values is necessary as it is explained in § 6. Nevertheless, the presented methodology remains always valid as with different numerical values the resulting feature orders may change but all the comprised procedures and computation steps remain valid.

6 Further Development

As previously discussed, the presented work consists a first step towards the definition of the feature importance for a given dataset. The developed methodology involves different schemas which need to be deeper exploited. Hence, multiple values of the involved hyperparameters should be considered (e.g. degree of polynomial features, values of the Ridge parameter α), various loss functions should be examined etc. In what follows, some points of interest regarding future work are introduced.

1. In § 4.2 the "kmeans++" clustering algorithm is exploited in order to classify data into distinct groups. However, in this work, data are also grouped according to "kmeans" and "DBSCAN" clustering algorithms. Consequently, the feature importance should be examined for each of these two grouping techniques. The required code for this work is available, but due to time constraints the implementations are not completed. The groups defined by the "kmeans" algorithm are available and saved in memory (see file cc_fi_3_1). Regarding the "DBSCAN" algorithm, the dbscan objects are saved in the computer memory. The necessary functions for defining the dataframes with the samples of each group are available but need to be called.

2. Other clustering algorithms should be considered and compare their results to the experimented ones. Some suggestions for clustering algorithms are: Affinity Propagation, Agglomerative Clustering, BIRCH, Mini-Batch K-Means, OPTICS, Mixture of Gaussians etc.

3. The considered loss function measuring the model performance is the "mean squared error". Additional loss functions should be tried such as "Huber Loss", "Mean squared logarithmic error", "Mean absolute error" etc.

4. During the data preparation, two datasets are created. The present work exploits only the one dataset. The second dataset should also be examined and provide a comparative study of the resulting outcomes.

5. An important issue to be highlighted is that all models are tried in a part of the available data (training set comprised of $48,803$ allocating the 20% of the data to the validation set). As "k-fold cross validation" technique is employed (k=5) the entire dataset should be provided to the model during the learning stage. Due to computer limitations (especially as other heavy programming tasks had to be executed simultaneously) implementations running on the entire dataset were very slow and deadlines could not be met. Consequently the train set is only used (involving 80% of the data).

6. A comparative study involving the best models resulting from each one of the three approaches could be established. For this task, each approach should be exhaustively explored for providing an optimized model regarding the employed dataset and the involved hyperparameter values.

7. Additional techniques on feature importance should be tried such as ones based on decision trees etc.

8. Development of different techniques regarding the treatment of missing values such as creation new AI models able to estimate new reliable values.

9. In order to achieve a clearer understanding of the feature importance, for each considered approach, the percentage of contribution of each feature to the price definition coud be resented. A simple calculation is only required of that using the provided feature coefficients.

10. Aiming at a deeper analysis of the considered problem, additional collected data providing information on the client characteristics could be useful to be employed.