Hadzhiyski Aleksandar 01426981
Holzner Peter Anton 01426733
Leitner Alexander 01525882

Machine Learning

Exercise 0 "Dataset description"

March 2020

- Classification Dataset - Caravana: Don't get Kicked https://www.openml.org/d/41162

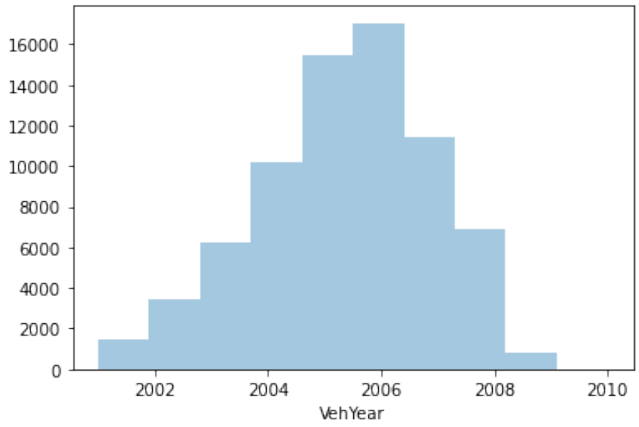
The dataset chosen by our group for classification contains data concerning a rather common issue in the USA - namely, the challenge to predict whether the purchase of a car from an auction is going to be a Kick (Bad buy) or not. The analysis in this case is aimed towards the position of a dealership rather than a private person buying a car however the overall approach stays the same in both cases.
There are many aspects to be considered when trying to model the factors which come into play when analyzing whether a car is a good purchase or not but the dataset tries to cover the main ones and in our opinion would prove useful.

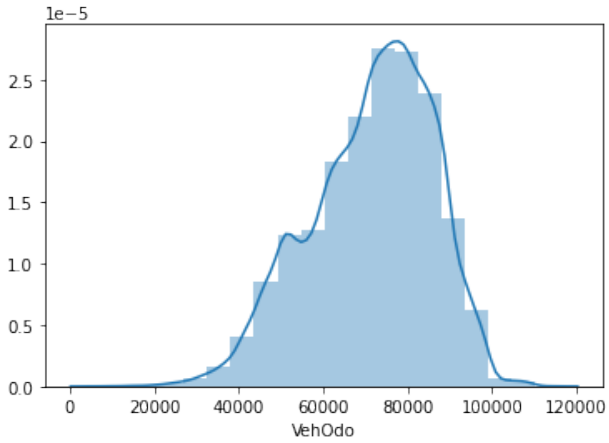
To keep our two datasets diverse, we have chosen this dataset to have the following features (which we have interpreted accordingly):

a. Number of samples - approximately 73000(a large number of samples)
b. Number of dimensions - 33 (relatively high dimensional)
c. Missing values - between 0 (0%) and 3200 (4.4%) spread across different attributes, ignoring the two attributes with 70000 missing values.

Our end goal is to be able to correctly predict a target attribute IsBadBuy which classifies whether a car is potentially a good or bad purchase (kick) based on its features. Thus our target attribute is a boolean value that states: True or False.

The values in the dataset are mostly nominal whereas some of the nominal categorical values simply represent a True/False relationship (Ex. IsOnlineSale column).
The more important numeric values such as vehicle age, vehicle odometer, are mostly normally distributed which helps show that the data can be considered a good example of real-world situations. (Plots on next page)
The categorical data, on the other hand is, quite diverse and nominal and in some cases could be considered also as ambiguously ordinal (Ex. Size column).





- Regression Dataset - Moneyball https://www.openml.org/d/41021

The second dataset we chose is based on a real-life story which was also depicted in the movie "Moneyball". The basic idea behind the dataset is that a well-analyzed batch of data could prove to be groundbreaking in some cases. The presumption is that, normally, in baseball, the scouts are the ones looking for new players to perfectly match them to a team in order to improve the team based on their knowledge and "feeling" for the game. At some point, a young statistics graduate meets the general manager of a baseball team and using only a similar dataset (without deep baseball knowledge) improves the team's performance immensely. This was the reason why we chose this dataset for these exercises.

In comparison this dataset is a bit smaller than the first one with the following properties:

a. Number of samples - approximately 1230 (a small number of samples)
b. Number of dimensions - 15 (low dimensional)
c. Few missing values

Our idea is to be able to correctly predict the target attribute Wins which, at the end of the season is the most important thing for any sports team. The distribution of the attribute is plotted at the end.

The nominal values within "Moneyball" are categorical, boolean and numeric: Team, League, Playoffs, RankSeason, RankPlayoffs and G, Year, W.
Some of these are obvious (like Team, League and so on). Others are not that obvious and therefore we have to describe them such as Runs Scored (RS), Runs Allowed (RA), Wins (W), On-Base Percentage (OBP), Slugging Percentage (SLG), Batting Average (BA), Playoffs (binary), RankSeason, RankPlayoffs, Games Played (G), Opponent On-Base Percentage (OOBP), Opponent Slugging Percentage (OSLG).
In this dataset there are only a few columns with missing values and these are also not important for the test data for the future predictions (Playoffs, RankSeasons and RankPlayoffs).

The most numbers of missing values are in the columns Opponent On-Base Percentage(OOBP) and Opponent Slugging Percentage(OSLG). The missing values from these attributes appear only in the earlier years. Maybe this value was not defined in that year and with that information these missing values are not important.