# Capstone Proposal

Lucas Saffi

Machine Learning Engineer Nanodegree

March 18th, 2019

## Project Proposal

### Domain Background

Pricing the most appropriate value for a product is a challenge for any company, because of the many market variables and its costs. Depending on the market this challenge can be even bigger because of the complexity, and certainly automotive segment is one of them.

Although technologies of vehicle sharing like Uber and Didi are becoming more common, the number of units of vehicles sold has grown continuously. According to Technavio, until 2021 the CAGR of used vehicles in the world will be 7%, reaching the number of 128 million of units. Considering this information and market competition, only prepared dealerships will excel and thrive. Therefore, it is crucial for them understand the market behavior to make better decisions, specially about price, which is fundamental to guarantee market volume and a good margin.

According to Listiani, in the research paper "Support Vector Regression Analysis for Price Prediction in a Car Leasing", the used car segment in Germany in 2004 generated 52 billion revenue and yields just 0.03 billion (1%), although it had a profit potential of 3%. In the past, it was hard to have a hardware good enough to compute fast and enable machine learning applications to effectively predict which price should be applied to maximize result, but today it is easy.

Vehicle pricing depends on factors like car model, brand, mileage, fuel type, power and time. Price correctly can help a dealership sell faster, with less cost, and with a more predictable margin. So, predict more accurately the most likely price of a car in market is certainly a solution with big impact for dealerships. Another reason for my interest about this project is that I am currently working in a dealership network, having knowledge of this can differentiate me to propose projects to achieve better and faster results for usage cars.

### Problem Statements

The problem to be solved is the price prediction of used automotive vehicles, which is a regression problem. Therefore, the model has the objective to determine the price a car will have based on past car ads samples. For example, if a car is a Volkswagen, model Golf, manual, 128 HP, regular gasoline and is going to be sold during a certain period, it is likely that will cost x dollars, that is what de model should predict. The price range is between € 0 and € 35.900 (removing absurd values).

## Datasets and Inputs

The dataset was offered on Kaggle with the name “Used cars database” and generated by Orges Leka, which collected 370.000 samples from ebay. It has the following information: date that the ads was created and last seen on ebay, mileage, type of seller, year of registration, offer type, gearbox, power, price, postal code, if vehicle is from control or test group (AB test) and if it was repaired or not.

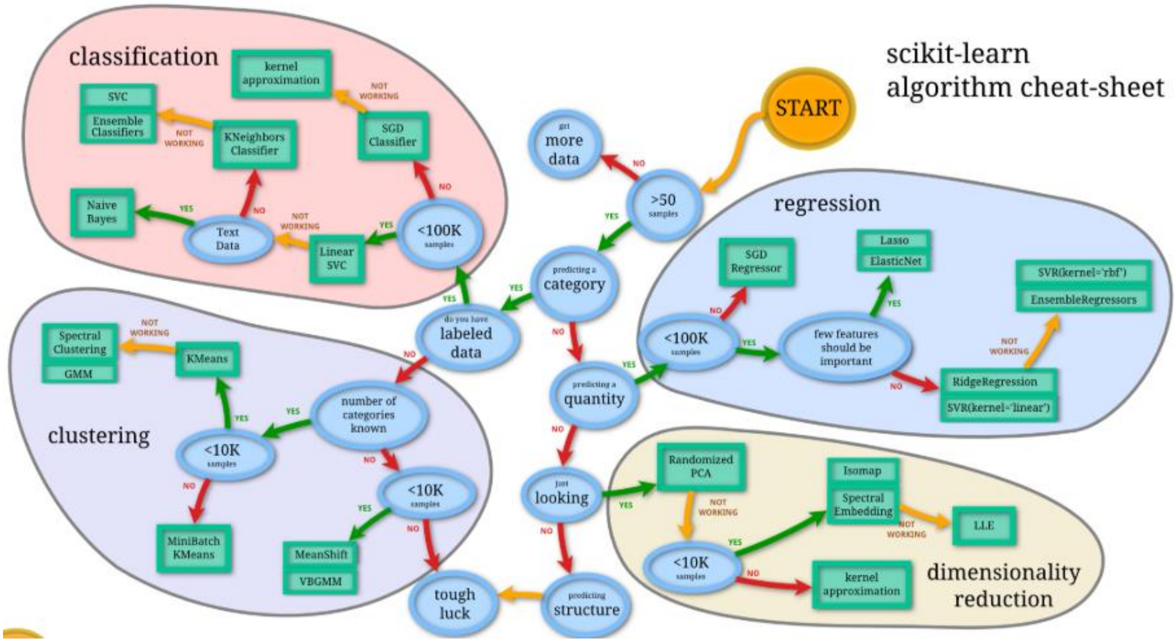
The link to access the dataset is: https://www.kaggle.com/orgesleka/used-cars-database.

## Solution Statement

My solution will use a Support Vector Regression (SVR) with different kernels: ‘linear’, ‘poly’, ‘rbf’ and a Random Forest Regressor. After each model being trained and predicted, the model with the best results will be tuned with different parameters setting. At the end, it will be compared to the benchmark, which is a Stochastic Gradient Descent Regressor (SGD Regressor) due to this type of model be highly recommended for big datasets.

## Benchmark Model

The benchmark model will be a Stochastic Gradient Descent regressor (SGD regressor) due to the number of dataset samples (370.000). This recommendation was taken based on Sklearn cheat sheet:



## Evaluation Metrics

The evaluation metrics used will be the score of each model, which is the $R^2$ of the prediction. This metric was chosen because measures how well the dependent variables explain the independent variable variance. So, it will measure how well future samples are likely to be predicted by the model. This metric is composed by the correct value price, prediction, mean and number of samples. The equation of this metric is:

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=0}^{n_{\text{samples}}-1} (y_i - \hat{y}_i)^2}{\sum_{i=0}^{n_{\text{samples}}-1} (y_i - \bar{y})^2}$$

## Project Design

My project will be divided in the following steps:

1. Preprocessing data: Since data is not perfectly separated it is necessary to remove values which does not make sense or are outliers, generate others from the current ones, for example, the number of days a vehicle took to be sold is not explicit, but can be determinated using data created and last seen columns values. Normalize data and transform categorical features to numbers are also included in this first step.
2. Data visualization: Before effectively train and predict, it is important to visualize how each feature relates with vehicle price, that is, the correlation between them and understand what is the most and less common values for each feature. For that a correlation map and histograms will be used.
3. Models training and prediction: 3 models of SVR with different kernels will be created (linear, poly, rbf), an ensemble model, Random Forest Regressor, and the benchmark, Stochastic Gradient Descent Regressor. Each of them is going to be trained and predict with standard parameters.
4. Model tuning: The model with best score value in the previous step, except the benchmark, will be selected and be tuned with GridSearchCV, to find the parameters which maximize score.
5. Comparison and conclusions: At the end, the final scores from benchmark and model tuned will be compared and conclusions about model quality predicting price of used vehicles will be made.

## References

1. www.scikit-learn.org/stable/modules/model_evaluation.html
2. www.scikit-learn.org/stable/tutorial/machine_learning_map
3. www.kaggle.com/orgesleka/used-cars-database
4. Listiani, Mariana. Support Vector Regression Analysis for Price Prediction in a Car Leasing Application. 2009.
5. medium.com/usf-msds/choosing-the-right-metric-for-machine-learning-models-part-1-a99d7d7414e4
6. www.globalfleet.com/en/news/global-used-car-market-grow-7-year-until-2021?a=FJA05&t%5B0%5D=Auto%20Remarketing&curl=1