**Statement of Work**

**Umutcan Asutlu**

**100797771**

**Background**

As artificial intelligence and machine learning solutions started to be used in different areas among many different industries, a used car retailor in the United States wants to use artificial intelligence technologies to increase their profit. They have offices in many locations where they buy and sell used cars. They are planning to buy cars in areas where they are sold for cheaper prices and sell those cars where they are more popular and worth more money.

**Objective**

The purpose of this project is maximizing the profit of this car retailor by using data analysis methodologies and machine learning solutions. Machine learning is one of the most used techniques for optimization for many different fields. The project will detect low price cars cities and make predictions for maximizing profit when selling them.

**Scope of Work**

With the help of the python this project will analyze the data that have been fed and will create many different visualizations to understand relations and answer many business questions but displaying trends based on geographic location. The project will detect low price cars in cities according to these trends to minimize company expenses. make predictions for different cities where to sell car for top price to maximize the profit. With same methods the project also keeps track of cities where cars are sold for high price. Using the data that generated by these methods and machine learning regression models the project can predict where certain type of car can be sold with maximum profit to the company. It will take role on both sales and acquisition.

Although there is an unknown transportation cost that may be a constraint for the project, it would still be insightful and increase the overall profit of the company.

**Data**

**Requirements**

The required data is data of used cars with all attributes that affect the price of the car is sold. The price that the car is sold with location that the car is sold are key factors for the project.

**Related Source**

The data is obtained from Kaggle (see reference 1). The data generated from Craigs List, widely used general purpose advertisement website from job to housing to cars. It consists of 25 columns which are from left to right: id, url, region, region\_url, price, year, manufacturer, model, condition, cylinders, fuel, odometer, title\_status, transmission, vin, drive, size, type, paint\_color, image\_url, description, county, state, lat, long. It includes all attributes that are needed to make predictions and it contains geographical location and price. There are total of 423,868 entries in the raw data. Most of the cars are made after 2003. There are a lot of null values in data for certain properties. According the attribute itself that is missing, the value will be replaced, or the entry will be removed.

Condition of the car is one of the most important attributes for algorithm that will work for benefit of retail company. The missing values on this column are all removed which was %58 of the existing data plus missing model values leads to total of 174,861 rows that are viable. 'id', 'url', 'region\_url', 'image\_url', 'description', 'county', 'vin' columns are also removed for convenience. There are total of 45 car manufacturers.

**Data Assumptions / Constraints**

The one big data concern is that there are a lot of car models due to every package for the car and attributes. They are important for defining the price so the information must be saved due to robust predictions about the price. But also it should be reshaped and grouped up in a convenient way that, same models with slight differences should not count as completely unique car models. There are total of 19272 unique models.

It is a very nice data contains a lot of features to feed on machine learning algorithm.

The data can be supported with external data sources if needed. There are multiple open data sources available online. Although there are no direct developer API of craigslist to gather the data but some data can be also gathered through RSS feed.

**Tasks**

* Data needs to be understood well
* Data must be cleaned, all useless information must be removed, needed missing values must be filled. Exploratory data analysis must be done.
* Correlations and patterns must be defined. Relation maps and geo maps must be generated.
* Different machine learning algorithms must be evaluated to pick best one to solve the problem.
* Implement the evaluated model for prediction results.
* Design a cloud-based solution to host the python codes and the results.
* Deploy all python codes and machine learning models to the cloud solution to finish implementation
* Test and debug

**Deliverables**

All deliverables will be delivered through GitHub.

* Statement of Work
* Used car data
* Python code for data clean
* Python code that given input, generates visualization for data insights
* Python machine learning code makes prediction based on generated data
* A platform for using the project

**Schedule**

Statement of Work 11/06/2020

Understanding data

Data Acquisition and Gathering Results 12/1/2020

Development Process

Project Delivery 12/18/2020

**Testing Process**

The project will be capable of detecting low prices and due to amount of existing data, it can generate accurate predictions with powerful machine learning algorithms. Tests will be done as algorithm will pick a car to buy and then show a place to sell with predicted price. The amount will be full profit for the retailor.

**References**

[1] <https://www.kaggle.com/austinreese/craigslist-carstrucks-data>