**Bahria University,**

**Karachi Campus**

****

**COURSE: SEN-322 SOFTWARE DESIGN AND ARCHITECTURE**

**PROJECT REPORT**

**SUPERVISED MACHINE LEARNING**

**USING LINEAR REGRESSION ALGORITHM**

**NAME REGISTRATION NO**

**ILSA ASIF 46107**

**ANEELA KARIM 46088**

**BSE-6B**

**ABSTRACT**

The majority of practical machine learning uses supervised learning. Supervised learning is where you have input variables (x) and an output variable (Y) and you use an algorithm to learn the mapping function from the input to the output.

Y = f(X)

The goal is to approximate the mapping function so well that when you have new input data (x) that you can predict the output variables (Y) for that data. It is called supervised learning because the process of algorithm learning from the training dataset can be thought of as a teacher supervising the learning process. We know the correct answers; the algorithm iteratively makes predictions on the training data and is corrected by the teacher. Learning stops when the algorithm achieves an acceptable level of performance.

Supervised learning problems can be further grouped into regression and classification problems.

Table of Contents

[1. Introduction 1](#_Toc9991262)

[2. Literature Review 1](#_Toc9991263)

[3. Methodology 1](#_Toc9991264)

[*2.1* *Dataset* 1](#_Toc9991265)

[*2.2* *Model / Algorithm* 2](#_Toc9991266)

[4. Results and Discussion 2](#_Toc9991267)

[5. Conclusion 2](#_Toc9991268)

[*References* 3](#_Toc9991269)

# Introduction

First we tried using the Google Trends API, since we thought if a user wanted to buy a car, they would definitely Google it on the web. However this API only gave us relative popularity scores out of a hundred. So comparing several thousand cars to each other was difficult. Furthermore, the Google Trends API had a query limit of 200 queries per day and 5 per hour.

We decided to switch the way we obtained popularity to using the Twitter streaming API. The streaming API obtains current tweets related to a search query. The streaming API was run for around 4 hours and we obtained 54,000 separate tweets. Our original plan was to use the Car Model and Make as the search queries, but with thousands of different Car Models and makes, the twitter streaming API was very slow to obtain data. We let it run for a day, and it only obtained a few hundred tweets. Hence we used the car brands as the search queries for these tweets.

# Literature Review

In Supervised Learning, algorithms learn from labeled data. After understanding the data, the algorithm determines which label should be given to new data based on pattern and associating the patterns to the unlabeled new data.

Supervised Learning can be divided into 2 categories i.e.

Classification & Regression

# 3. Methodology

The classification script is written in Python, making use of the Jet Brains Pycharm. And we used scikit-learn’s classification functions.

First of all we read the dataset using pandas. Then divide the whole dataset into 75% as training set of (x, y) and 25% as testing set (x, y). And train the model Decision Tree Classifier with training set (x, y) and checking the accuracy by using testing set. And then predicting the any single random sample of input (the desired car model).

We used the label encoding to encode the each feature of every parameter into a particular index (numbered form).

## *Dataset*

The car feature data was collected using the Edmunds car API. A script was written in Node.js using to scrape car details. Edmunds API contains various vehicle specs, consumer and critic ratings, and prices. We obtained 11,915 car samples. Some of the cars were used, and some were new. New cars used their average MSRP as the price, while old cars used their TMV value as the price.

## *Model / Algorithm*

“Linear regression model”

# 4. Results and Discussion

1. **Popularity (target):**

The top ten most popular cars are as follows, from most to least: Ford, BMW, Audi, Ferrari, Honda, Toyota, Nissan, Dodge, Kia, and Porsche.

2. **Important Features:**

The most important features in determining the price of cars are found to be the engine horsepower, the engine fuel type, and the engine cylinder, which is sensual since they determine how well a car runs. Model types, the make, and the year come after; these have more to do with the brand and superficial appeal of the car, but are still deciding factors nonetheless.

3. **Most Overpriced Cars and Brands:**

We hoped to do a more in-depth analysis of car prices and the effects of branding. We have seen that a car’s make and model type may affect its price quite significantly. In order to determine is a car is overpriced or not, we ran the classifier without these features. The most overpriced cars are found to be the ones with the greatest difference between its projected price (by the algorithm) and its actual price (the testing values). In addition to individual models, we took a look at the most overpriced brands.

4. **Accuracy Achievement:**

Our designed algorithm is about (86 to 90)% accurate regarding popularity prediction.

5. **Final Achievement:**

Our final achievement is to calculate the average popularity (that how much car will be sold ) of a new predicted model that is going to be launched by the particular company.

# 5. Conclusion

The results do not disappoint; indeed it seems that the most overpriced brands are those that sell ridiculously priced cars that everyone wants to buy but don’t have the money for it.

# *References*

<https://www.kaggle.com/jshih7/car-price-prediction>

<https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/>

<https://towardsdatascience.com/supervised-machine-learning-classification-5e685fe18a6d>