Machine Learning Nanodegree

Capstone proposal

Name: Abdullah Hany

Email: abdullah.hany22@gmail.com

Domain background

Supervised learning is one of the most promising field in Machine-learning where already much development has taken place and is currently used in real world application.

Supervised learning is a Machine-learning task of teaching a function to map an input to an output based on previous input-output pairs

Nowadays almost everyone has a car, people tend to change their cars frequently in order to enjoy up to date technology, since cars are not cheap, they tend to sell their old car to gain some money and use it to buy a new car

An example of an academic paper where machine learning is applied to this type of problem is https://www.researchgate.net/publication/319306871_Predicting_the_Price_of_Used_Cars_using_Machine_Learning_Techniques

The source of data that I will be using is https://www.kaggle.com/orgesleka/used-cars-database#autos.csv

Problem statement

When people tend to sell their cars they either give it to a dealership or post it online, however sometimes people ask for twice the money of that car is worth or sometimes sellers are afraid of being scammed either by the dealership or by the buyer

Therefore, this regression project will help in providing an adequate price for the used cars based on the market.

User should provide same basic information about the car as the input for the program, for example:

- Make and model of the car
- Type of the vehicle
- Year of registration
- Gear box

The output of project will be a prediction of the car price.

Datasets and inputs

The used cars dataset includes Over 370000 used cars scraped with Scrapy from Ebay-Kleinanzeigen.

The dataset contains

- name : "name" of the car
- price : the price on the ad to sell the car
- vehicleType
- yearOfRegistration : at which year the car was first registered
- gearbox
- monthOfRegistration : at which month the car was first registered
- fuelType
- brand

The goal of the program is to correctly identify the price of the car i.e. the price column

Some normalization will be done to insure that the data are not skewed then, I will split the data into 80% training set and 20% testing set using train_test_split() function.

Solution statement

I thought of making a Machine-learning algorithm using the linear regression technique, train it using the mentioned dataset.

After training, the algorithm will be able to understand the relationship between all the columns we mentioned above in the dataset section and the price of the car and will be able to predict the price of the car for the user, based on the market value.

Benchmark

As this is a Kaggle dataset, a benchmark model would be the best Kaggle score which, comes at 82% accuracy score using the RandomForestRegressor model

Evaluation metrics

The model prediction for this model can be evaluated in several ways.

Since the evaluation technique used by most kernels in the score function embedded in the regression model, however I will be using both the score function embedded in the regression model and R2 score to get an accurate difference between predicted labels and actual labels

Project design

My method would be using classic linear regression models.

Steps of designing the project

- read the data
- drop some useless columns based on observing their values like, seller and nrOfPictures columns have only one value and I do not think that postal code , dateCreated nor dateCrawled are relevant for price prediction
- clean the data by removing some NULL values and duplicates
- use some visualization to better understand
- if the data is skewed and not balanced , some normalization techniques can be used like natural log or Standardization scaling
- convert categorical features to numbers
- split the data into training and testing sets
- XGBoost, LightGBM, SVM and random forest regression models can be used
- Grid search technique can be used to help in parameter tuning

Tools and Libraries used: Python, Jupiter Notebook, pandas, numpy, scikit learn, seaborn and matplotlib.

Other libraries will be added if necessary.

References

- https://www.kaggle.com/orgesleka/used-cars-database/kernels
- https://en.wikipedia.org/wiki/Supervised_learning
- https://scikit-learn.org/stable/modules/generated/sklearn.metrics.r2_score.html
- Supervised learning section in Udacity Machine-learning Nanodegree program