Machine Learning Engineer Nanodegree

1 CAPSTONE PROPOSAL

Charles Frederic Atienza

February 27, 2020

2 PROPOSAL

2.1 DOMAIN BACKGROUND

Supervised learning is rapidly becoming one of the most researched and utilized field of study in Machine learning in commerce. It is being used from all industries from agriculture, marketing, and even stock trading. Given the versatility of supervised learning, its application to markets outside of business-to-consumer transactions is worth exploring. Consumer-to-consumer marketplaces like Craigslist and Carousell have its own economy which is driven by the consumers themselves. Supervised learning could potentially standardize these economies so that each consumer has a sufficient understanding of what kind of factors contribute to the prices of goods on these marketplaces.

My interest on these marketplaces is mainly used cars. Used cars, compared to other goods, have a lot of contributing factors that determine their price like model, mileage, and condition. Having bought used cars myself, I found it difficult to gauge how much each factor contributes to its price on the market.

2.2 PROBLEM STATEMENT

Often on consumer-to-consumer marketplaces, sellers tend to base their items' price off other items of the same kind's price. Sellers of used cars, however, tend to come up with their own price based on estimates on various factors like the model, mileage, and condition of the car. There is also the presence of car flippers who buy used cars on below market price and sell on a considerable profit margin that does not always translate to how much the car is actually worth. With supervised learning and ample data, we can determine a reasonable price for a used car given its many quantifiable specifications.

2.3 DATASETS AND INPUTS

The dataset is taken from the Used Cars Dataset on Kaggle which is licensed under public domain. Its content consists of every active used vehicle posting within the United States on the Craigslist as of January 2020. The dataset contains only one file *vehicles.csv* that contains the entries.

INPUT DATA FEATURES

- *id* - entry ID
- *url* - listing URL

- *region* - Craigslist region
- *region_url* - region URL
- *price* - entry price
- *year* - entry year
- *manufacturer* - manufacturer of vehicle
- *model* - model of vehicle
- *condition* - condition of vehicle
- *cylinders* - number of cylinders
- *fuel* - fuel type
- *odometer* - miles travelled by vehicle
- *title_status* - title status of vehicle
- *transmission* - transmission of vehicle
- *vin* - vehicle identification number
- *drive* - type of drive
- *size* - size of vehicle
- *type* - generic type of vehicle
- *paint_color* - color of vehicle
- *image_url* - image URL
- *description* - listed description of vehicle
- *county* - useless column left in by mistake
- *state* - state of listing
- *lat* - latitude of listing
- *long* - longitude of listing

This dataset will be considered labelled data since the car's *price* is among the data features. Major features of the data that will be examined such as *manufacturer, model, odometer,* and *transmission*. Along with these, secondary features' correlation to the price will also be examined, namely, *fuel, drive, type,* and *paint color*. The *condition* feature will also be investigated thoroughly if any patterns emerge from it seeing as almost half of the dataset has no entry for it.

2.4 Solution Statement

Of the dataset, only entries from January 2015 to December 2019 will be used. Of these entries, 50% will be used for training, 25% for validation, and the remaining 25% for testing. Input features will be filtered into only those that have significant correlation to the *price.* These input features will be decided upon after data exploration. The model will use Amazon SageMaker's built-in *LinearLearner* algorithm with a regressor predictor type in order for the model to produce a quantitative output. The output will be a single integer whose value represents the model's prediction of the input used car's price based on its features.

2.5 Benchmark Model

Given that the model's output is quantitative, two other models whose output is quantitative will be fed the same dataset in order to benchmark the original *LinearLearner* model's performance. The first will be another of Amazon SageMaker's built-in algorithm, *XGBoost* with a linear regression objective. The other will be a custom *PyTorch* model with a linear output layer and Mean Squared Error as its loss function.

2.6 Evaluation Metrics

Each model's performance will be evaluated based on its output's accuracy. Which is, the average of each output's percentage distance to the car entry's label.

$$accuracy = \frac{\sum_{i=1}^{n} 1 - \frac{|x_i - y_i|}{y_i}}{n}$$

2.7 Project Design

2.7.1 Data Exploration

The dataset will be loaded in and read. Of the loaded data, only entries from January 2015 to December 2019 will be taken for data exploration.

Through the most appropriate plots, relation of the following features to *price* will be explored:

- *manufacturer*
- *model*
- *odometer*
- *mileage*
- *fuel*
- *drive*
- *type*
- *paint color*

It will also be explored if the value of *condition* produces any pattern in relation to *price*. The reason for this is the assumption that sellers whose used cars' condition is below good are inclined to not specify the condition at all in fear of scaring off potential buyers.

One other thing that is worth exploring is that if the elapsed time from the entry's year (given by *year*) to the car model year (given in *model*) on its own has significant impact on the *price*.

2.7.2 Data Pre-processing

Categorical features of the dataset will be one-hot encoded while numerical features will be normalized around 0.

If the correlation between the elapsed time from the *year* to the *model*'s year to *price* proves promising, the elapsed time will be put into a new column *elapsed_years* and normalized as well.

2.7.3 Feature Selection

Each feature's correlation to each other will be calculated and the features whose correlations to others are below 0.9 will be selected as the input features for the model.

2.7.4 Models Creation

The input data will be shuffled and separated into a training set, validation set, and testing set. All the models will be trained through SageMaker and will employ hyperparameter tuning to get the best performance possible.

2.7.4.1 Linear Learner

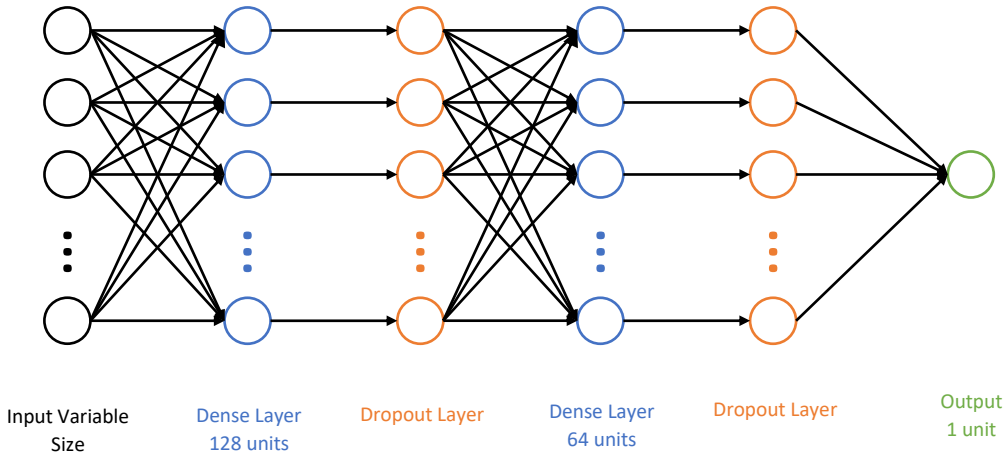
An estimator for Amazon SageMaker's *LinearLearner* algorithm will be fitted with the training set. Its predictor will be set to regression in order for it to output a quantitative output. The fitting will use the validation set for validation and then, the testing set for performance and accuracy evaluation.

2.7.4.2 XGBoost

An estimator for Amazon SageMaker's *XGBoost* algorithm will be fitted with the training set. Its objective will be set to logistic regression in order for it to output a quantitative output. The fitting will use the validation set for validation and then, the testing set for performance and accuracy evaluation.

2.7.4.3 Custom PyTorch Model

A custom neural network will be created with the PyTorch framework. It will have the below structure.



Same as the *LinearLearner* and *XGBoost* models, The fitting will use the validation set for validation and then, the testing set for performance and accuracy evaluation.

2.8 REFERENCES

2.8.1 Hudgeon, D., & Nichol, R. (2020). Machine learning for business: using Amazon SageMaker and Jupyter. Retrieved from https://docs.aws.amazon.com/sagemaker/latest/dg/xgboost.html

2.8.2 Hudgeon, D., & Nichol, R. (2020). Machine learning for business: using Amazon SageMaker and Jupyter. Retrieved from https://docs.aws.amazon.com/sagemaker/latest/dg/linear-learner.html

2.8.3 Reese, A. (2020, January 7). Used Cars Dataset. Retrieved from https://www.kaggle.com/austinreese/craigslist-carstrucks-data