Machine Learning Engineer Nanodegree

Capstone Project Proposal

Domain Background

Online shopping has a vital role in our daily lives due to its low cost, highly convenience, ease of use, and so forth. Selling an item involves a trade-off between higher price and shorter selling time. One area of research that has received little attention is the time-series variation of the relative pricing and how the relative product pricing changes over time.

The term time series refers to a series of data-points that indexed in time order, in other words, a sequence of observations taken sequentially in time. Thus, time series forecasting is the process of using a model to predict future values based on previously observed values. Data used for time series prediction is non-stationary e.g. economic, weather, stock price, and retail sales.

Most existing price prediction models for selling second hand items only focus on selling items in specific domain. For example, several work have studied car price prediction [1-3] and predicted the end price of online auctions for laptop and PDA categories [4, 5]. While these efforts tried to make the best deal based on individual profit maximization, no model is designed with the aim of predicting an adequate price for a product. Furthermore, most price prediction models rely only on limited number of features or online product reviews/descriptions; they don't consider other main features like different prices of different product categories.

For time series analysis, a recent research utilized the sentiment analysis of news, txt, with the time series analysis to forecast the e-commerce prices [6]. The proposed work is evaluated for only Apple mobile phone and the forecasted prices are for a new product not second hand products.

Also, there is a related work that uses dataset consists of both images and contextual information to predict the houses prices, "House price estimation from visual and textual features", https://arxiv.org/pdf/1609.08399.pdf

Problem Statement

The output of the time series prediction problems is a scalar value, which is applicable in problems such as weather or stock prices. But in some cases, it is required not to predict a single value. For instance, predicting the price of second-hand product at a certain time point is a range, depending on the quality of the used product. Thus, the time series model should predict a range instead of a scalar. In this proposal, I am looking to adapt machine learning times series models to predict a range of values instead of a single value at a given time point.

Datasets and Inputs

To predict a range at a given time point, I used a dataset that contains ads of second-hand products. Each second-hand item has a description, an image, a product type, and a price. There are different categories of the products, the list of products is in ***Figure 1***.

product type	No_of_observations
furniture	11795
electronics	8992
generalforsale	7776
sportinggoods	6822
car & truck	6477
householditems	6392
autoparts	6331
musicalinstruments	5646
bicycles	5560
motorcycles/scooters	5527
baby & kidstuff	5427
appliances	5400
clothing & accessories	4601
antiques	4543
collectibles	3881
tools	2784
computers	2567
arts & crafts	2035
toys & games	2033
photo/video	1651
computerparts	1085
motorcycleparts	1008
farm & garden	997
cellphones	836

Figure 1 Dataset: list of product types with the number of ads at each type

We used a **price indices** for different product types to mimic the price change of products through time**,** which is a normalized average of price relatives for a given class of goods or services during a given interval of time. Depend on this standard price index, we changed the ads prices for each product of 210 time point. Then, we distributed that product type ads (items) over the time points, such that each time point has at least 20 ads, see **Figure 2**. Each time point has a minimum price and maximum price. We changed the ads' prices to follow the price index curve so that the prices have a hidden equation. This hidden equation will be the metric to decide the degree the model successes. In other words, fitting the price index curve is the model target.

Once the range is predicted for a given product at a given time point, the image and the text of the ad will be used to locate the predicted price of this second-hand product. The text and image of the ad at hand will be evaluated by another ML model to determine its position in the predicted range; this should measure the quality of the second-hand product. For example, if we want to predict a price range of a certain product after a month, the first model will predict a price range, say 100$ to 125$. Then, the second model will use the ad text and image to find a quality measure between 0 and 1. If the result of the second model is 0.75. Thus, the final predicted price of this second-hand product is 100 + 25*0.75 = 118.75$, where 100 is the min price and 25 is the range value.

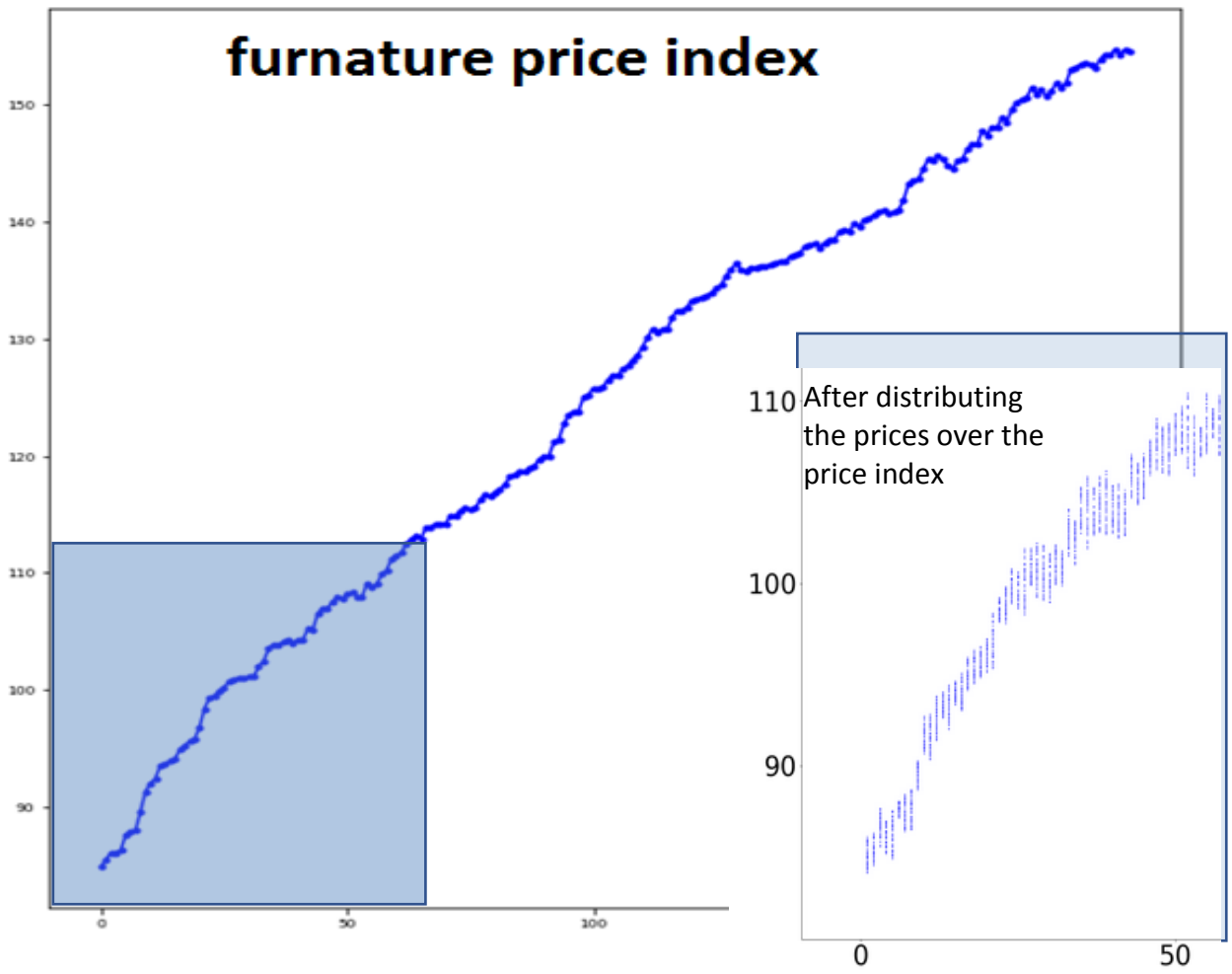

Figure 2 Price Index

Solution Statement

In this problem, we want to answer two main questions.

First, as the used products prices falls in different ranges over different time points, then what are the predicted price ranges for the sold items in the future?

To answer this question:

1- We suggest to use two models one to predict the minimum price and another model to predict the maximum price, by this we have an interval of the predicted range of prices of each product type.
2- Another solution is to use a model to predict the range value (difference between maximum and minimum), with another model that predict the minimum price. Then we have a range of values (minimum + range value).

Second, the text and image of the ad at hand will be evaluated by another ML model (linear regression or logistic regression) to determine its position in the predicted range; this should measure the quality of the second-hand product.

Notice:

One additional feature that can be added to the project: if the customer wants to sell a product and he want to put in online within some period e.g. a month. Then we will recommend him a day to put the product online.

Benchmark Model

To the best of our knowledge, we did not find a machine learning model that can predict a range of values. But predicting prices with text and images dataset is applicable by using Neural Networks or Support Vector Regression. Thus, we will compare one of these solutions with our work.

Evaluation Metric

The problem being addressed is divided into two sub-problems:

1. **Time series problem**:

predicting price range for the expected prices following time points in future. we changed the ads' prices to follow the price index curve so that the prices have a hidden equation. This hidden equation will be the metric to decide the degree the model successes. Our target is to fit this curve using our machine learning model.

This problem will be handled with these methods

- Deep learning method, where our evaluation metric is MSE.
- Regression method, where our evaluation metric is R-squared.
- Time series statistical model ARIMA, where our evaluation metric is MSE.

And finally, state the method with the best results

2. Regression problem: predicting a specific price in the predicted range resulting from the time series problem, which will be our model quality indicator. As regression problem, I suggest to use R-squared metric as an evaluation metric for this problem.

Project Design

The project will go as follow:

- We will organize the dataset according to price index and clean it from null or duplicate ads.
- Using the standard price index for each product type and draw a curve of it, which will be our indicator of model accuracy.
- Based on solution section,
  For quality level, we will use linear regression or logistic regression as it is a simple regression problem.
  For the time series part, we will evaluate the following three models
  1- Regression model, linear regression.
  2- Deep learning model, CNN or RNN.
  3- Time series statistical model, ARIMA.
- Then we will use the price index equation to measure our model accuracy.
- Repeat the previous process till find the best accuracy.

1. Pal, N., et al., *How much is my car worth? A methodology for predicting used cars prices using Random Forest.* arXiv preprint arXiv:1711.06970, 2017.
2. Zhang, X., Z. Zhang, and C. Qiu, *Model of Predicting the Price Range of Used Car.* 2017.
3. Noor, K. and S. Jan, *Vehicle Price Prediction System using Machine Learning Techniques.* International Journal of Computer Applications, 2017. **167**(9).
4. Ghani, R. *Price prediction and insurance for online auctions*. in *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*. 2005. ACM.
5. Raykhel, I. and D. Ventura. *Real-time Automatic Price Prediction for eBay Online Trading*. in *IAAI*. 2009.
6. Kuo-Kun Tseng, R.F.-Y.L., Hongfu Zhou, Kevin Jati Kurniajaya, Qianyu Li, *Price prediction of e-commerce products through Internet sentiment analysis.* Electron Commer Res, 2017. **October**(10).