Capstone Proposal

Sergio Castillo

January 21st, 2018

Proposal

DomainBackground

E-commerce has changed the manner in how people buy and sell products. In the past, a person had to go to the store, chose the desired item and pay. Nowadays, consumers do not need to physically inspect the product to acquire it. Instead, they have developed patterns based on the information that retailers provide on their website[1]. This data has helped sellers to understand their client's behaviour and use it for different purposes. Such as, forecasting sales[2], customer's segmentation[3], detect trending items [4], and products' price prediction[5][6][7] which is the area that this study will cover.

Price prediction has been a subject of study for a long time in the economy, specifically in the stock market[8]. Different techniques have been proposed, from probabilistic methods[9] in the 70's to ensemble techniques combining neural networks(NN) with swarm optimization[10], or deep learning [11]. Currently, research to predict prices has been adopted in more areas as, travel[12][13], real estate[14][15], agriculture[16], or retail[1][2][5][6]. In this research, we will focus only in the retail industry.

The importance of predicting prices in the e-commerce retail industry can be divided into 3 areas: 1) for the customer, 2) for the retailer, and 3) goods manufacturers. First, the customer can obtain a baseline price or an expected price for a certain item[17]. Second, the seller can use the prediction to define a price when introducing a new product, or forecast sales[2]. Last, the manufacturer can use the information for a deeper study that could include the total number of items to be sold, and define a strategy in their production line to estimate the number of raw materials needed for the expected demand[18].

The objective of this study is to have hands-on experience in a dataset provided by Mercari, one of the leading retail companies in Japan via a Kaggle competition. Also, as part of the contest, and the motivation to earn a monetary prize, for me it is important to compare my models with the best people in the data science community that can only be found in Kaggle.

Problem Statement

From Kaggle[19]:

It can be hard to know how much something's really worth. Small details can mean big differences in pricing.For instance, a sweater with almost the same description can vary from $355.00 to $9.99.

Product pricing gets even harder at scale, considering just how many products are sold online. Clothing has strong seasonal pricing trends and is heavily influenced by brand names, while electronics have fluctuating prices based on product specs.

Mercari, Japan's biggest community-powered shopping app, knows this problem deeply. They'd like to offer pricing suggestions to sellers, but this is tough because their sellers are enabled to put just about anything, or any bundle of things, on Mercari's marketplace.

In this competition, Mercari's challenging you to build an algorithm that automatically suggests the right product prices. You'll be provided user-inputted text descriptions of their products, including details like product category name, brand name, and item condition.

Datasets & Inputs

The dataset is provided by Mercari. It contains two files, one for training and one for testing.

The training dataset has 1482535 instances with 7 features and one target value, described as following:

- **train_id** or **test_id** - the id of the listing
- **name** - the title of the listing. Note that we have cleaned the data to remove text that look like prices (e.g. $20) to avoid leakage. These removed prices are represented as [rm]
- **item_condition_id** - the condition of the items provided by the seller
- **category_name** - category of the listing
- **brand_name**
- **price** - the price that the item was sold for. This is the target variable that you will predict. The unit is USD. This column doesn't exist in test.tsv since that is what you will predict.
- **shipping** - 1 if shipping fee is paid by seller and 0 by buyer
- **item_description** - the full description of the item. Note that we have cleaned the data to remove text that look like prices (e.g. $20) to avoid leakage. These removed prices are represented as [rm]

Solution Statement

The price prediction problem can be solved by developing regression models. As input, we have the dataset discussed in the previous section. The output is the predicted price for each item. There are different approaches we can adopt as support vector machine(SVM), recurrent neural networks(RNN), or linear regressions. This will be discussed in detail later. One of the most interesting challenges in this project is the pre-processing part in the feature called ”item_description” which should be attacked using natural language processing(NLP).

Benchmark

Based on the literature found, the techniques that can be implemented with reliable results are 3: Decision trees regression, linear regression[6], and recurrent neural networks(RNN)[1]. Besides this supervised learning algorithms, I would like to try support vector regressions(SVR). These models have to be compared based on the evaluation metric, which is the Root Mean Squared Logarithmic Error(RMSLE), explained in the next section.

Evaluation Metrics

(approx. 1-2 paragraphs)

The evaluation metric is the **Root Mean Squared Logarithmic Error**. Which is calculated as follows:

$$\varepsilon = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(log(p_i+1)-log(a_i+1))^2}$$

Where:

- ε is the RMSLE value (score)
- n is the total number of observations in the (public/private) data set.
- p_i is your prediction of price
- a_i is the actual sale price for i
- $log(x)$ is the natural logarithm of x

Project Design

This project can be divided in 7 steps:

1. **Exploration**.- It focuses on analyzing each feature, their correlation, and distribution. It is a descriptive analysis of the entire data. Graphically explain how the dataset is composed.

2. **Pre-processing**.-Overall, the test dataset contains missing values, we need to define what to do with those values, replace them?, use filling methods?. Also, we have to tokenize 3 features: category_name, brand_name, item_description. Part of the challenge as mentioned previously is implementing NLP in the feature called item_description. This might affect directly in the models' performance. It might be useful to implement dimension reduction. For instance, Lasso or PCA.

3. **Feature selection**.- After we pre-process the data, we need to identify which features could help the model to predict better prices. This step is highly linked to the pre-processing since we need to be very careful with the tokenization.

4. **Develop models**.- In this stage, the techniques mentioned in the Solution Statement are going to be implemented.

5. **Compare results**.- After running an recording the performance of each model, an analysis of each one is going to be conducted. Part of this analysis is going to be later used in the conclusions.

6. **Tunning model(s)**.- If, after having the results two models have similar performance, we would keep both and tune them in order to reduce the RMSLE. Otherwise, if just one model provides the best results, select that one so it could improve the price prediction.

7. **Conclusion**.- At this point, we summarize the entire process, provide lessons learned, the best results, areas of improvement, limitations during the project, and future work.

Bibliography

[1] Reid Pryzant, Young-joo Chung, and Dan Jurafsky. “Predicting Sales from the Language of Product Descriptions”. In: (2017).

[2] Kris Johnson Ferreira, Bin Hong Alex Lee, and David Simchi-Levi. “Analytics for an online retailer: Demand forecasting and price optimization”. In: *Manufacturing & Service Operations Management* 18.1 (2015), pp. 69–88.

[3] Dipanjan Sarkar, Raghav Bali, and Tushar Sharma. “Customer Segmentation and Effective Cross Selling”. In: *Practical Machine Learning with Python*. Springer, 2018, pp. 373–405.

[4] KM Anil Kumar et al. “Effective Approaches for Classification and Rating of Users Reviews”. In: *Proceedings of International Conference on Cognition and Recognition*. Springer. 2018, pp. 1–9.

[5] Hassan Waqar Ahmad. “Prediction of retail prices using local competitors”. PhD thesis. Faculty of Graduate Studies and Research, University of Regina, 2014.

[6] Prajakta Badhe. “Retail pricing prediction using linear regression”. In: *Neural Networks & Machine Learning* 1.1 (2017), pp. 1–1.

[7] Michael P Wellman, Eric Sodomka, and Amy Greenwald. “Self-confirming price prediction strategies for simultaneous one-shot auctions”. In: *arXiv preprint arXiv:1210.4915* (2012).

[8] Jerry Felsen. “Learning pattern recognition techniques applied to stock market forecasting”. In: *IEEE Transactions on Systems, Man, and Cybernetics* 5.6 (1975), pp. 583–594.

[9] Manak C Gupta. “Money Supply and Stock Prices: A Probabilistic Approach”. In: *Journal of Financial and Quantitative Analysis* 9.1 (1974), pp. 57–68.

[10] Salim Lahmiri. “A Technical Analysis Information Fusion Approach for Stock Price Analysis and Modeling”. In: *Fluctuation and Noise Letters* (2018), p. 1850007.

[11] Xiao Ding et al. “Deep Learning for Event-Driven Stock Prediction.” In: *Ijcai*. 2015, pp. 2327–2333.

[12] Slava Kisilevich, Daniel Keim, and Lior Rokach. “A GIS-based decision support system for hotel room rate estimation and temporal price prediction: The hotel brokers’ context”. In: *Decision Support Systems* 54.2 (2013), pp. 1119–1133.

[13] Stacey Mumbower, Laurie A Garrow, and Matthew J Higgins. “Estimating flight-level price elasticities using online airline data: A first step toward integrating pricing, demand, and revenue optimization”. In: *Transportation Research Part A: Policy and Practice* 66 (2014), pp. 196–212.

[14] EJTG van der Burgt. “Data Engineering for house price prediction”. In: (2017).

[15] Adyan Nur Alfiyatin et al. "Modeling House Price Prediction using Regression Analysis and Particle Swarm Optimization". In: ().

[16] Changshou Luo et al. "Prediction of vegetable price based on Neural Network and Genetic Algorithm". In: *International Conference on Computer and Computing Technologies in Agriculture*. Springer. 2010, pp. 672–681.

[17] Austan D Goolsbee and Peter J Klenow. "Internet Rising, Prices Falling: Measuring Inflation in a World of E-Commerce". In: ().

[18] VL Raju Chinthalapati, Narahari Yadati, and Ravikumar Karumanchi. "Learning dynamic prices in multiseller electronic retail markets with price sensitive customers, stochastic demands, and inventory replenishments". In: *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 36.1 (2006), pp. 92–106.

[19] *Mercari Price Suggestion Challenge Description*. `https://www.kaggle.com/c/mercari-price-suggestion-challenge`. Accessed: 2018-01-21.