Machine Learning Engineer Nanodegree

Price Prediction for ecommerce products

MANCEÑIDO, AGUSTIN
Febrauray 14th, 2019

Domain Background

Brand manufacturers should constantly look to evaluate their pricing, learn what works, and make improvements. If the price is too low, you run the risk of not being able to obtain a healthy profit. On the other hand, if the price is too high and you run the risk of not making any sale. eCommerce companies requires an holistic understanding on pricing, and should be positioned as a marketing weapon and a conversion rate optimizer [1].

Optimal pricing is as much art as it is science, and it can mean the difference between an avalanche of orders flowing out or a mountain of excess stock sitting somewhere. Pricing products and services online is one of the most exciting and complex exercises and, as an online business, finding the right formula is one of the most important questions to solve [2].

Problem Statement

Given a dataset of historical sells corresponding to a product category of an e-commerce site, it is necessary to develop an algorithm to predict the optimal competitive price for a new product to be published. The algorithm needs to determine what is the relevant data in order to get the best prediction for the price taking into account seasonal prices, market fluctuations, economic trends, better sellers, etc.

Datasets and Inputs

The dataset to be used on this project has been collected from an Argentine ecommerce site [3] using the official API [4]. It consists 795295 records of sells with with 8 features each, all taken between 03/17/2017 and 05/10/2018. The following features are included:

- **title**: name of the product provided by the seller.
- **sellerid**: unique id for each seller.
- **state / city**: seller location.
- **pos**: position where the product was found in the search.
- **sqty**: number of products selled.
- **price**: price of the product.
- **created_at**: timestamp for the sell.

Solution Statement

The dataset will be analyzed to understand how each feature is related to the others, consider the importance of each feature and find out which are actually relevant for understanding the price of the product. Then the data will preprocess to create a better representation of price by performing a scaling and detecting (and removing if needed) outliers to obtain significant and meaningful results and get the prices related to the best sellers. Finally, a Neural Network model will be constructed and trained to be able to predict the price using Keras [5] implementation of Tensorflow [6] library.

Benchmark Model

This project will use XGboost [7] as its primary benchmark. XGboost is a quick and efficient and implements machine learning algorithms under the Gradient Boosting framework that is easy to use and is also enormously flexible. Over the past year and half 50% of the Kaggle competitions have used XGboost as a key part of the winning solution according to Ben Hamner, CTO of Kaggle [8].

Evaluation Metrics

The measure of performance for this project will be the Mean squared logarithmic error (MSLE) [9] between the predicted and actual values of the product price for a test subset of values extracted from the dataset. This metric penalizes an under-predicted estimate greater than an over-predicted estimate, which is appropriate for a price prediction that will be used by the sellers.

Project Design

The theoretical workflow for approaching a solution will be in the following order:

- Data exploration
    - Loading libraries and dataset
    - Dimensions of the data
    - Statistical analysis
    - Features correlation
- Data preprocessing
    - Preprocess feature columns
    - Data cleaning
    - Feature engineering
    - Scaling / Normalizing data
    - Training, Validation and Test split
- Evaluate models
    - Prepare benchmark model
    - Train and test benchmark model
    - Create Neural network model
    - Train and test model
- Model tuning and compare results
- Conclusion

References

[1] Lemonstand Blog: 5 eCommerce Pricing Strategies to Help You Achieve Profitable Growth.
[2] Lemonstand Blog: 5 Pricing Strategies You Should Test On Your eCommerce Store.
[3] Mercado Libre, Inc. (https://www.mercadolibre.com.ar/)
[4] Mercado Libre API. (https://developers.mercadolibre.com.ar/)
[5] Keras: A high-level neural networks API. (https://keras.io/)
[6] TensorFlow: OS library for high performance numerical computation. (https://www.tensorflow.org/)
[7] XGBoost: An optimized distributed gradient boosting library. (https://xgboost.ai/)
[8] Udacity Interview with Ben Hammer.
[9] scikit-learn: Mean squared logarithmic error (https://scikit-learn.org/stable/modules/model_evaluation.html#mean-squared-logarithmic-error).