Machine Learning Engineer Nanodegree

Capstone Proposal

Yingying Hu
September 20th, 2020

Proposal

Domain Background

"How well is a product likely to sell?" is a crucial question for every retailer. Before the coming of eCommerce, brick-and-mortar businesses can only rely on intuition and experience when forecasting customer's demand on one product. Nowadays, more and more people choose to shop online because of the convenience and product varieties. At the same time, business owners can gain more customer data by tracking their activity on the shopping website. Analyzing and understanding what data suggests product sales performance will help businesses to optimize their stocks. Furthermore, a machine learning algorithm can bring it to the next level by predicting product sales and make decisions on product stock for business owners.

Problem Statement

The goal of this project is to create a predictive model to estimate the product sales.
Tasks including:

1. Explore and preprocess the “Sales of summer clothes in E-commerce Wish” dataset on Kaggle. Wish is a popular American online e-commerce platform that facilitates transactions between sellers and buyers.
2. Select important features for predicting sales for products
3. Choose an ML model and train the model
4. Check the model accuracy by using Mean Accuracy and Root Mean Square Error (rmse)

Datasets and Inputs

The “Sales of summer clothes in E-commerce Wish” dataset is public on Kaggle. The table is from crawling the result page after searching “summer” on Wish, and it was collected in August 2020. It contains information about summer clothes products and their sellers, such as the product id, retail price, ratings, seller’s reputation, etc. These usually drive customer’s decision on purchasing a product, so they can be relevant features on predicting a product sale.

The data is in a csv file, and it has 43 columns. Data cleaning will be needed since there are some duplicated entries and some missing values. Some columns, like product_id and product_url will not be able to use by the model, so they will be removed.

Solution Statement

In the project, an ML model will be built and trained on selected features from the Wish dataset. The model will be tested on a test set and its accuracy will be calculated using the Root Mean Square Error

Benchmark Model

The benchmark model can be a K-Nearest Neighbors with 10 clusters. This KNN model is shown in a public notebook shared by a Kaggle user. The model’s Mean Absolute Error is about 57.9

Evaluation Metrics

Since the goal is to predict product sales, the model’s label is numerical. The model’s evaluation metrics will be Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). MAE and RMSE are two common metrics used t measure accuracy for continuous variables. Since the benchmark model uses the MAE, it will be used for comparison purpose.

$$\mathrm{MAE} = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

Project Design

(approx. 1 page)

This project will be worked on a Jupyter notebook environment. The workflow will be:

1. Set up a git repository and a notebook instance on the AWS SageMaker platform
2. Download the data file from Kaggle and upload to a notebook instance
3. Explore, visualize, and clean the table, including handling missing values and duplicated entries
4. Feature engineering, such as check the correlation on numerical variables and preprocess the categorical features
5. Split the data into train, validation, and test set
6. Choose 2-3 base models, such as Random Forest, XGBoost, and Neural Networks.
7. Training the selected base models on the train set, and compare them and the benchmark model on the validation set using MAE and RMSE
8. Select the best performance model and do hyperparameter tunning
9. Check the final model accuracy on the test set and get the MAE and RMSE