YAP 470 Project - Predict Future Sales

Alp Bora Kirte
TOBB Economy and Technology University
akirte@etu.edu.tr, Ankara/Turkey
https://github.com/AlpBora/YAP_470_Project-Predict_Future_Sales.git

***Abstract*—Prediction applications are quite common in machine learning. That’s why for this project i have participated a Kaggle competition named "Predict Future Sales". In this paper i will describe the problem, overall plan for approaching the problem and offer some possible solutions .**

***Index Terms*—forecasting, time-series data, future sale prediction**

I. Introduction

One of the largest Russian software firms - 1C Company provided challenging time-series dataset consisting of daily historical sales data. The task is to forecast the total amount of products sold in every shop for the test set. My goal is to predict total sales for every product and store in the next month.

II. Problem Statement

The dataset to be used is consisting of 6 csv files; **sales_train.csv**(the training set, daily historical data from January 2013 to October 2015.), **test.csv** (the test set), **sample_submission.csv** (a sample submission file in the correct format), **items.csv** (supplemental information about the items/products), **item_categories.csv** (supplemental information about the items categories), **shops.csv** (supplemental information about the shops).

The dataset has 11 data fields; **ID** (an Id that represents a (Shop, Item) tuple within the test set), **shop id** (unique identifier of a shop), **item_id** (unique identifier of a product), **item_category_id** (unique identifier of item category), **item_cnt_day** (number of products sold), **item price** (current price of an item), **date** (date in format dd/mm/yyyy), **date_block_num** (a consecutive month number, used for convenience, January 2013 is 0, February 2013 is 1,..., October 2015 is 33), **item_name** (name of item), **shop_name** (name of shop), **item_category_name** (name of item category).

Because, for the reason that dataset contains time series data, i must think of other machine learning solutions and use other algorithms. For example now rows are independent so, i can no longer split the train and test data randomly and can not use K-Fold cross validation. This will result in data leakage beetwen past and future datas so i have used "sklearn.model_selection.TimeSeriesSplit " method for cross validation.

III. Technical Approach

I have not yer determined the exact model to use but from my research **LGBM, CatBoost, XGboost** models are commonly used for this competition. Furthermore, Simple Moving Average (**SMA**), Exponential Smoothing (**SES**), Autoregressive Integration Moving Average (**ARIMA**), Neural Network (**NN**) and **Croston** are one of the effective statistical techniques available for time series forecasting ones. I will also try **KNN** and **Decision Tree** with few of these proposed techniques and decide which one is fastest and more accurate.

Aside from machine learning models, i will also need a feature extraction method. I will use window, static or lag features (adding previous sale data as features) for this. Calculating the autocorrelation beetwen months with lag also can be usefull, though i am going to need to make the time series stationary before calculating correlation.

Some features are not necessary i droped them and also ’date’ column isn’t in right form so i have added **’month’ ’year’** and **’date_of_weak’** columns to **’sales_train.csv’**

Furthermore using some libraries like tsfresh, feature-engine, darts and sktime can also be usefull since i am dealing with forcasting instead of tabular regression. Sktime allows it’s users to solve forecasting problems using machine learning models from scikit-learn so i will be using it’s functions.

IV. Intermediate/Preliminary Results:

I have examined the data in Exploratory Data Analysis part and here are the some conclusions i have drawn;

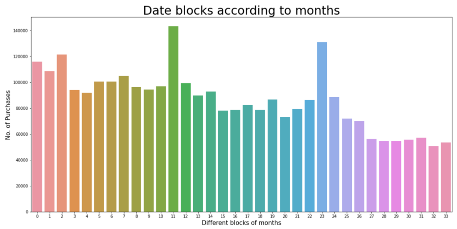
Some shops don’t have data for more than a year. November and December are missing in 2015.

There are some stores that did not exist in 2013 and some closed in 3 years duration.

In **’sales_train.csv’** the **’item_cnt_day’** column has -1 as values in some days. I have learned from a kaggle discussion that this means buyer has returned that item back.

Some items are no longer sold and some new items started to be sold.

The number of purchases are the highest in 11th and 23rd month.



I have also analysed the busiest day, month and year and while the most busiest year is 2013, the less one is 2015.