Machine Learning Engineer Nanodegree

Capstone Proposal

Zijun Hu Udacity
Oktober 7th, 2020

Rossmann Store Sales

1. Domain Background

For capstone project I‘ll get into the thema data mining. This Project will be built to conduct the data from Featured Prediction Competition „Rossmann Store Sales“ hosted at Kaggle. Rossmann operates over 3,000 drug stores in 7 European countries. Currently, Rossmann store managers are tasked with predicting their daily sales for up to six weeks in advance. Store sales are influenced by many factors, including promotions, competition, school and state holidays, seasonality, and locality. With thousands of individual managers predicting sales based on their unique circumstances, the accuracy of results can be quite varied.

Through this work I could have a thorough glance in the occupational area data mining in which I‘ll occupy.

2. Problem Statement

This project is a forecasting problem using time series. The goal is to predict the next 6 weeks of sales for 1,115 stores located across Germany using models trained by a time series from the span 2013-1-1 to 2015-7-31. The accuracy and quality of the modeling will be evaluated on the Root Mean Squre Percentage Error.

3. Datasets and Inputs

The sales of a store are impacted by many features, such as store type, competitions, promotions, holidays, day of week ans so on. A time series with some features will be used to train the models and then make the predicting of the Sales for the next 6 weeks.

Some relevant features are listed below:

- **Customers** - the number of customers on a given day
- **Open** - an indicator for whether the store was open: 0 = closed, 1 = open
- **StateHoliday** - indicates a state holiday. Normally all stores, with few exceptions, are closed on state holidays. Note that all schools are closed on public holidays and weekends. a = public holiday, b = Easter holiday, c = Christmas, 0 = None
- **SchoolHoliday** - indicates if the (Store, Date) was affected by the closure of public schools
- **StoreType** - differentiates between 4 different store models: a, b, c, d
- **Assortment** - describes an assortment level: a = basic, b = extra, c = extended
- **CompetitionDistance** - distance in meters to the nearest competitor store
- **CompetitionOpenSince[Month/Year]** - gives the approximate year and month of the time the nearest competitor was opened

- **Promo** - indicates whether a store is running a promo on that day
- **Promo2** - Promo2 is a continuing and consecutive promotion for some stores: 0 = store is not participating, 1 = store is participating
- **Promo2Since[Year/Week]** - describes the year and calendar week when the store started participating in Promo2
- **PromoInterval** - describes the consecutive intervals Promo2 is started, naming the months the promotion is started anew. E.g. "Feb,May,Aug,Nov" means each round starts in February, May, August, November of any given year for that store

4. Solution Statement

The features are the main factor to There are essentially two appoaches to solve this Problem. One is regression, e.g. Xgboost, random forest. The other one is time series forecast using such as ARIMA model.

Basically this project will go though with XGBoost and random forest regressor. Eventually also make the predict with a time series analysis. At the end a comparison among them will be conducted.

5. Benchmark Model

It is recomended to use the XGBoost due to its good regression performance in this project. XGBoost stands for „Extreme Gradient Boosting. It‘s an implementation of Gradient Boosted Decision trees designed for speed and performance.

This algorythm has these advantages:

- Regularization: XGboost uses a more regularized model formalization to control over-fitting.
- Automated missing values handling: XGB uses a "learned" default direction for the missing values. "Learned" means learned in the tree construction process by choosing the best direction that optimizes the training loss.
- Interactive feature analysis (yet implemented only in R): plots the structure of decision trees with splits and leaves.
- Feature importance analysis: a sorted barplot of the most significant variables.

Random Forest will be also used to solve the predicting problem to conduct a comparison.

6. Evaluation Metrics

The Forecast values are valuated on the Root Mean Square Percentage Error (RMSPE). The RMSPE is calculated as

$$\text{RMSPE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\hat{y}_i}{y_i}\right)^2}$$

where y_i denotes the sales of a single store on a single day and yhat_i denotes the corresponding prediction. Any day and store with 0 sales is ignored in scoring.

The deduced result should be camparable with the outcomes in Leaderboard.

7. Project Design

1) Data cleaning

What are the data types? Are there NaN‘s, how are the NaN‘s distributed and how to deal with the NaN‘s, just drop them or replace them with some other values?

2) Data exploration and analysis

Plot the trends of Sales respectively according to month, day of week, store type and assortment to have a thorough overview of the Sales so that we can understand the data better.

3) Feature Engineering

In oder to produce a forecasting with high qulaity. We must produce the appropriate features derived from the current features that could best characterise the Sales. I‘ll do some analysis of the chosen features by plotting some correlation heatmap about them.

4) Model training

The train data will be splitted to train and validation Set. Basically two models will be applied to do the training, Random Forest and XGBoost. They are both the appoach of regression. Eventually I‘ll also apply the method of time Series Analysis.

5) Model Parameter tuning

With the the weight correction the the parameter of Model XGBoost and Random Forest will be tuned.

6) Predicting

7) Discussion

In the end the performance of them would be compared and discussed.