

Capstone Proposal

Rossmann Store Sales Prediction



March 30, 2018

Written by: Mu(Marvin) Yuan

# Domain Background

The project of Kaggle competition was chosen for the final project of Udacity Machine Learning Engineer Nanodegree. The Kaggle competition is about a chain drug store named Rossmann which is the second largest drug store chain with over 3600 retailer stores in Europe, which is founded in 1972[1]. This competition is to predict the drug store’s 6 weeks ahead daily sales for 1115 Germany stores.

The motivation for investigating this forecast sales project is because that this is a very good opportunity for me to do a Kaggle project with guided. Kaggle is a huge platform for me to get in the field of machine learning with data scientists from all over the world. By familiar with Kaggle competition, a larger network as well as knowledge will be built.

# Problem Statement

The problem is given 15 different features to predict one of the 15, which is Sales. From the data given, it is a supervised learning problem since the output (sales) is given. The inputs are in different types, such as numeric variables and categorical variables. Numeric variables included continuous variable such as data, and discrete variable such as sales. Categorical variable included ordinal variable such as assortment and nominal variable such as promo to indicate if a specific store running a promo that day. The metric for success is via Kaggle’s leaderboard which is based on accuracy and so on.

# Datasets and Inputs

The given data are formed from 15 different features as shown in Table 1.

Table : Drug store sale prediction feature list [2]

|  |  |
| --- | --- |
| Feature Names | Description |
| ID | an Id that represents a (Store, Date) duple within the test set |
| Store | a unique Id for each store |
| Sales | the turnover for any given day |
| 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 |

The output should be sales which should be predicted from remaining 14 input features.

# Solution Statement

Clearly, the Rossmann sales prediction is a structural supervised problem. The solution should be written as a python code file with supervised learning algorithm. Due to research, one of the potential algorithm can be Xgboost, although the other algorithms will be used for training as well. The result will be shown in Kaggle leaderboard and accuracy.

# Benchmark Model

A benchmark model supposed to be fast and high performance. By measurable terms, the total training time won’t last over 5 minutes, and the evaluation score should be better than 80%. A benchmark model will be evaluated from several different supervised learning algorithms.

# Evaluation Metrics

After the benchmark model chosen, the parameter set is the key to optimize the training performance. The F-beta score will be used to evaluation the training score:

Equation : F-beta score equation



Accuracy measures how often the classifier makes the correct prediction. It’s the ratio of the number of correct predictions to the total number of predictions (the number of test data points).

Precision tells us what proportion of messages we classified as spam, were spam. It is a ratio of true positives (words classified as spam, and which are spam) to all positives (all words classified as spam, irrespective of whether that was the correct classification), in other words it is the ratio of

Recall(sensitivity) tells us what proportion of messages that were spam were classified by us as spam. It is a ratio of true positives (words classified as spam, and which are spam) to all the words that were actually spam, in other words it is the ratio of

# Project Design

The general structure will be data exploration from raw data to get a general idea of the data structures. Next step is data wrangling, also known as data preprocess, several different steps will be proceeded including normalization, data transformation, shuffle split data and so on. After the data was preprocessed, it will be used for training. In the training stage, several algorithms will be chosen and evaluated by its training time and result. One model will be selected for final benchmark model. After the benchmark model chosen, the parameter set is the key to optimize the training performance. By tuning the parameters, the optimized model creates.

# Reference

[1] <https://en.wikipedia.org/wiki/Rossmann_(company)>

[2] https://www.kaggle.com/c/rossmann-store-sales/data