Machine Learning Engineer Nanodegree

Capstone Proposal

Jiawei Jin
December 31st, 2018

Proposal

Domain Background

Rossmann operates 3,000 pharmacies in seven countries in Europe. Currently, managers at the Rossmann store are required to predict their daily sales for the next six weeks. Store sales are subject to many factors, including promotions, competition, school holidays and legal holidays. As thousands of operators predict sales based on their unique circumstances, the accuracy of the results can vary widely. In this project, it will predict the daily sales of 1,115 Rossmann stores in Germany for 6 weeks. Reliable sales forecasts will allow store operators to increase productivity and enthusiasm to create more efficient staffing arrangements. By helping Rossmann create a strong predictive model, I'll help operators stay focused on what matters to them: their customers and their teams.

Basically, this is a data mining challenge and data mining is the process of discovering patterns in large data sets involving methods of machine learning (ACM SIGKDD., 2006). Therefore, I want to use machine learning methods to solve this problem.

Problem Statement

I want to set up machine learning on old sales records and then predict the daily sales after six weeks. The specific steps and possible problems are as follows:

• Clean and organize the data. Possible problems are the handling of outliers and the handling of missing values.

• Perform a single variable analysis on the data, analyze multiple variables, and extract high-correlation variables that affect the results. I may encounter problems with how to extract highly correlated eigenvalues.

• Modeling with XGBoost, how to adjust the data of the training model to suit the input requirements of the XGBoost model is the challenge.

• How to find the most suitable super ginseng.

• Predict and evaluate the results. Using RMSPE as the evaluation which requires a visualization of the results and the need to supplement relevant knowledge.

Datasets and Inputs

I am provided with historical sales data for 1,115 Rossmann stores which has already been split into training set (1.02 million rows) and testing set (41.1 thousand rows) by the organization.

Most of the fields are self-explanatory. The following are descriptions for those that aren't:

- 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 (this is what you are predicting)
- 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

Solution Statement

- View feature types and characteristics of missing values, and supplement media values or set upper and lower limits to generate randomly. Find and drop out outliers based on the distribution of individual features.
- Sort the correlation of eigenvalues to select features that have a positive effect.
- Cut data into Train and Test and perform matrix to meet XGBoost input requirements

- Find the most appropriate parameters by continually trying and using the advanced usage of the XGBoost algorithm.
- Show the relevant visual results to complete the submission of the final result.

Benchmark Model

For the forecast of sales, the artificial error rate is assumed to be 10%. I use some other models, such as the GB model, to compare the performance of XGBoost.

Evaluation Metrics

I use RMSPE as a validation function, the lower the value, the smaller the difference. It is a measure of the difference between the predicted value of the model and the actual observed value.

Project Design

- Clean and organize the data, process missing values, unify data types, split part features, merge parts, and remove some features.
- Combine store with train and test data to eliminate and split some features.
- Visually analyze data variables and remove outliers.
- Combine the Train and Store data into train and test splits and matrix to prepare for training.
- Configure the xgboost parameters and start the first training.
- Summarize the results of the first training, further optimize the features, and further optimize the parameters
- Start the second training and find the best super ginseng by configuring the xgboost parameters
- Record the results of the training assessment and save the model.
- Use the model to make predictions on the test data and save the results as a submission.

Reference

"Data Mining Curriculum". ACM SIGKDD. 2006-04-30. Retrieved 2018-12-27.