Machine Learning Engineer Nanodegree

Capstone Project

Predict how sales of weather-sensitive products are affected by snow and rain

Capstone Proposal

Domain Background

This project is to predict sales of weather-sensitive products around the time of major weather events. The features - sales data for 111 products whose sales may be affected by weather(milk, bread etc). These 111 products sold in 45 different Walmart locations and these locations are covered by 20 weather locations that provides weather data.

Regression analysis is a statistical process that estimates relationships among variables. It helps to understand the relationship between dependent (Y) and independent variables(X) and how value on dependent variable changes with change in independent variable. It identifies what independent variables are necessary for the value of dependent variable.

This project will use supervised learning of Regression analysis to build models on training data set and make prediction based on given test data. The test data contains stores, dates and item for which we need to predict “units” sold.

This project is proposed from Walmart’s competition at Kaggle : https://www.kaggle.com/c/walmart-recruiting-sales-in-stormy-weather. There are many weather related events, mostly snow, occurs in an Area where I live and this project will help me extend this knowledge in other areas - for example, How to estimate Salt for snow event? How much snow will fall based on weather related data.

Walmart operates 11,450 stores in 27 countries. Extreme weather events, like hurricanes, blizzards, and floods, can have a huge impact on sales at the store and product level. In this challenge Walmart is asking to predict sales of weather-sensitive products like milk, bread and umbrellas during the major weather event. This will help Walmart stores to correctly maintain the level of inventory and

avoid out-of-stock.

Field descriptions

- date - the day of sales or weather
- store_nbr - an id representing one of the 45 stores
- station_nbr - an id representing one of 20 weather stations
- item_nbr - an id representing one of the 111 products
- units - the quantity sold of an item on a given day
- id - a triplet representing a store_nbr, item_nbr, and date. Form the id by concatenating these (in that order) with an underscore. E.g. "2_1_2013-04-01" represents store 2, item 1, sold on 2013-04-01.

Problem Statement

This forecasting model will allow stores to predict sales of weather related items for upcoming weather events. This may help stores to maintain optimal inventory and increase in sale based on weather patterns.

Datasets and Inputs

The dataset is based on Kaggle competition and here are the list of dataset provided:

- key.csv - the relational mapping between stores and the weather stations that cover them
- sampleSubmission.csv - file that gives the prediction format
- train.csv - sales data for all stores & dates in the training set
- test.csv - stores & dates for forecasting (missing 'units', which you must predict) - password for the file: Work4WalmarT
- weather.csv - a file containing the NOAA weather information for each station and day
- noaa_weather_qclcd_documentation.pdf - a guide to understand the data provided in the weather.csv file

```
Training Data Size (4617600, 4)
Test Data Size (526917, 3)
Weather Data Size: (20517, 20)
```

Solution Statement

The solution is consists of three parts.

1. Prepare data for training - This will include joining test/train data with weather data, remove any null values, the Items codes are not same across 45 locations- come up some strategy with same code/Id. Converting and non-numeric data into numeric values.

2. This project will use set of Superwised regression machine learning algorithms and choose the one with the highest coefficient of correlation.

3. Once model is selected then we will make prediction with test data for our final report.

Benchmark Models

This project will use following benchmark algorithms.

- SciKit learn -Linear regression and Gradient Boosting Regressor ensemble algorithm.

Evaluation Metrics

The following metrics will be used for model evaluation and selection.

- coefficient of determination(model score)

- mean squared error

Project Design

This project is based on Kaggle competition offered at : https://www.kaggle.com/c/walmart-recruiting-sales-in-stormy-weather/data

Here is high level project design approach:

- This project will be built on Python 3.5 and will use libraries like sklearn, numpy,

pandas, matplotlib etc.

- Prepare data for training. This involves joining data from “train.csv”, “stores.csv” and “weather.csv.”

train_data = “train” data join with “stores” join with “weather” data
test_data = “test” data join with “stores” join with “weather” data

- Cleanup null values from training and test data.

- Drop features that exists in test data, but not in training data and vice-versa.

- There is no separate set of validation data provided so for cross-validate use 80/20 rules to split data into training(80%) and validation data(20%)

- Use Grid Search technique to tune hyper parameter, max_depth =1..10, for Gradient boosting Regressor.

- Compare r2 score and mean squared error for classifiers and select the best performer - high r2 score.

- Make prediction with selected classifier on test data.

Bibliography

1. Coefficient of determination - https://en.wikipedia.org/wiki/Coefficient_of_determination

2. Mean Squared Error - https://en.wikipedia.org/wiki/Mean_squared_error

3. Walmart Kaggle Link : https://www.kaggle.com/c/walmart-recruiting-sales-in-stormy-weather