Machine Learning Engineer Nanodegree - Capstone Proposal

You

March 8, 2018

1 Domain Background

This project provides a simple solution to Recruit Restaurant Visitor Forecasting on Kaggle. It is a public competition and availabel for everyone. Rather than focusing to become winner in the competition which requires good hardware and amount of time, this project aim to give walkthrough the process of using different approaches (Statistical and Machine Learning) to solve a time-series problems. This will get beginners to understand the basic knowledge of analysis time-series data and can help them to apply in the practice.

2 Problem Statement

Kaggle is a platform for data science competition, where many people trying to solve company's problem by building predictive modeling and produce the best model. This project is one of the kaggle competition that involve many Japanese restaurants and help them forecasting how many customers to expect each day in the future. The forecasting is necessary for restaurant's owner because this forecasting is an important aid the effectively and efficiently planning the schedule staff members and purchase ingredients. The forecasting won't be easy to make because many unexpected variables affect the visitor's judgment, for example, weather condition, preferences, date, popularity, etc. Some things are easier to forecast than others. How much data is available and how further are we are going to forecast will affect the forecasting model we build. It will be more difficult when the restaurant only has little data.

3 Datasets and Inputs

The dataset is already publicly accessible on Kaggle sites. It comes from many sources and need to be download separately (total: 71.3 MB). In summary, the data comes from two separate websites that collect user information:

- Hot Pepper Gourmet (hpg): Japan's gourmet site, here users can search restaurants and also make a reservation online on restaurant in japan.
- AirREGI / Restaurant Board (air): A point of sales system specialized for Restaurant, can be used as reservation control and cash register system.

4 Solution Statement

Our goal is to forecast of customers visitation each restaurant from AirRegi. In order to make a better forecast, understanding the historical pattern (past data) and features (cross-sectional data) are needed. We might assume that some aspects in the past will continue in the future and some information that we observed affect the forecast.

After we identify the pattern and features, we use them to create our model. We try to make models by using a different approach. In general, two approaches that we are going to use for forecasting visitors in all restaurants is:

- Statistical approach: Forecast per-restaurant only by using trend, seasonal and cyclic pattern in historical data. We can use univariate ARIMA for this kind of approach.
- Machine Learning approach: Forecast by using supervised learning with features like date-related, time-related, location-related, weather, etc.

5 Benchmark Model

In this challange, benchmark model is not provided. So in order to get comparison of our model, we need to create our own benchmark model. The benchmark that we are going to use in this project is untuned Random Forest model. Usually Random Forest tends to overfit the training data, it will create better prediction in training data but worst when use unseen data.

6 Evaluation Metrics

Every kaggle competition have different quality metrics that used by participants to evaluate their model performance. Restaurant Forecasting competition features Root Mean Squared Logarithmic Error (RMSLE). RMSLE and RMSE used to find out the difference between predicted values and the actual one. RMSLE usually used when we don't want to penalize huge differences in the predicted and actual values when both of them are huge numbers.

The RMSLE is calculated as:

$$\sqrt{\frac{1}{N}\sum_{i=1}^{n}(log(P_i+1)-log(\alpha_i+1))^2}$$

where:

- n is the total number of observations.
- P_i is your prediction of visitors.
- α_i is the actual number of visitors.
- $log(x)$ is the natural logarithm of x

7 Project Design

The workflow of this project will divided into 3 sections. The summary of each section provided below.

7.1 Data Exploration

Before making some predictions, we need to properly understand and get the insight from Restaurants Forecasting's dataset. By doing data exploration, we poke around the data and getting a sense of what happened in the data. Due to data comes from different sources, we might have to use query and merge data to get more insight. Also, to make a better model, it will be useful to get some basic statistics, plot the data, and understand the features that correlated to the customer visits.

7.2 Statistical Approach

To forecast something, we need to understand the several factors that affect it. A good forecasting captures the patterns and relationship between variables and historical data that impact future events, not only random fluctuation or noise. To understand the pattern and behaviors in time-series data, we need to split the data into several components and then analyze each component.

Most of the statistical approach only use historical data to predict the future. By identifying the historical pattern, we can reconstruct data and use it to forecast the future. ARIMA model able to explain the trend and seasonality pattern from the historical data and reconstructed it. It

is a popular and traditional method for forecasting time series data. Usually, when people come into time-series problem, this approach is preferable and already widely used.

We start with understanding the pattern from ACF and PACF plot. ACF and PACF is a tool that used to identify the historical pattern from time series, how much related is the data from its past. By understanding this plot, we can try to do a simple forecasting model. After that, we try to tune the parameters and adding some features to make better predictions.

7.3 Machine Learning Approach

Machine Learning can be used for forecasting a time series model. What makes time series different from normal regression is that they are time-dependent. The basic assumption of regression is that the observation is independent of each other. Time series data also have some non-linear pattern in it, for example, seasonal and trends. To make it into regression problem, we need to capture the patterns and include them in the feature.

XGBoost algorithm can be used to solve the problem with regression method. It is a popular algorithm and usually used for kaggle competition because of its flexibility and predictive power. The only problem to use XGBoost is that there are many hyper-parameters that can be tuned to improve the model. After we create a simple model, we will try to tune the hyper-parameters by using grid search to improve the model.