Capstone Proposal Lucas Kuttner Amin

Machine Learning Engineer

Nanodegree

April, 10, 2018

Sales Forecast prediction

Proposal

Companies often finds itself allocating too much resources on unnecessary segments or executing failed investments, declining sales and lost opportunities for promotions and product releases because they don't have the proper data and future predictability.

Sales forecasting allows companies to execute all of the tasks that need future data and actions that need predictability. It is the process of predicting what a certain team, salesperson or company will sell in any given future time period. They are used mainly by company managers and directors to anticipate decisions like hiring, resource management, budgeting and available working capital.

Companies usually build their sales forecast based on manually on excel or any other financial tool, through many hypothesis, analysis and even competition sales information, choosing between bottom-up or top-down approaches, based on their business plan.

The main proposal of this capstone project is to solve the sales forecasting problem through machine learning supervised learning techniques of predictive analysis, testing different models and comparing their results of train time, test time, mean square error and complexity involved.

This project has been retrieved from a Kaggle's website competition, here, as there is a lot of introductive materials and knowledge rich kernels to learn from. Plus the necessity of the main market to reliable sales forecast, I've chosen this problem to solve as capstone project.

Domain Background

Sales forecast is the process of predicting a company, team or individual sales for accurate decision-making and to identify fast and clearly low revenue periods, caused by unproductivity, market crysis or competitors advantage on aggressive sales approaches. There are many usual sales forecasting methods, such as: Historical-based, pipeline-based and opportunity stage.

Although these methods are vastly used, it takes a lot of time and crew to achieve an good sales forecast, with the risk of it being wrong on its prediction. This is why sales forecast should be tested and validated with machine learning. The usage of past data (even if it is from other companies) can train efficient models to predict future sales result for any given company, if not alone, to give support and confidence to a sales forecast team.

Problem Statement

Using machine learning to execute sales forecasting makes the process a lot faster and more reliable, even if it is used only as a support for the sales forecast team. This problem was proposed at Kaggle, at the competition “Predict Future Sales” here. The problem may be divided into several relevant features extracted from past data, measured its result based on success rate of regression and reproduced by any given notebook, with usage of same dataset used from the model.

Datasets and Inputs

The dataset is built from daily historical sales data, with item categories, shops and sales archives, each with its supportive file (Test data or supplemental information). The general information provided is helpful for predicting future sales, using item prices, quantity sold, date, categories and ids.

The dataset is provided by a large Russian software firm 1C Company, retrieved at the competition Predict Future Sales at Kaggle.com.

Solution Statement

The plan is to first implement a model using supervised learning random forests and then compare its results to other methods like RNN, with and without data processing. All the created models will run with the same data, comparing its results with root mean squared error (RMSE) at the end of the project.

Benchmark Model

The benchmark will consist on comparing the results of the used within the project and with all the competitors at the challenge “Predict future sales”, adding the results to train time, test time, and root mean squared error (RMSE).

Evaluation Metrics

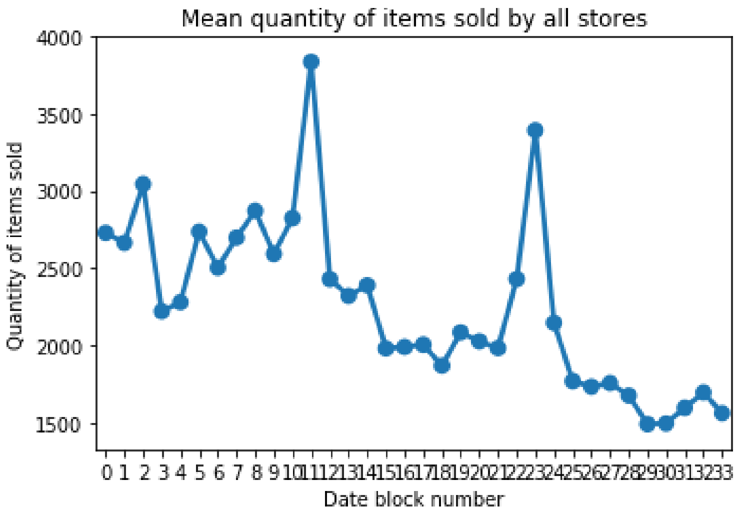
The evaluation metric will be the same as used at Kaggle’s competition; root mean squared error at the test set. This is the used evaluation because it measures precisely the average of the squares of the errors, which is the difference between the estimator and the result.

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{t=1}^{n} (\hat{y}_t - y_t)^2}{n}}.$$

Project Design

The approach will begin analyzing the data and extracting insights from it, using cumulative variance, mean values, std, analysis of the time-series behavior of all graphics. Second, investigating the results of a random forest model approach, setting up graphics and score information. Lastly, I'll explore and document a long term short memory neural network approach.

The data analysis consists of getting insights from graphical representations of data. The following graph contains the mean quantity of sales among all shops which have enough data points.



There is a clear pattern of descent, disregarding the peaks on the final months of each year of data. Using the overall information along with each individual category item seasonal sales as input data for the ensembles' ExtraTreesRegressor and after that, the LTSM recurrent neural network.

Using other available kernels at Kaggle's for learning and getting insights, there may be more models for comparing and data preprocessing.

Support kaggle kernels:

https://www.kaggle.com/the1owl/playing-in-the-sandbox
https://www.kaggle.com/carmnejsu/sales-forecast-lstm-67-beginner-friendly
https://www.kaggle.com/jagangupta/time-series-basics-exploring-traditional-ts

References

- https://blog.hubspot.com/sales/sales-forecasting
- https://trackmaven.com/marketing-dictionary/sales-forecasting/
- https://www.thebusinessplanshop.com/blog/en/entry/how_to_forecast_sales
- https://machinelearningmastery.com/challenging-machine-learning-time-series-forecasting-problems/
- https://www.kdnuggets.com/2017/05/springml-sales-forecasting-using-machine-learning.html
- https://rstudio-pubs-static.s3.amazonaws.com/105869_f6e7f8d4e0434c40bd939a3d1e792af9.html
- http://www.insightsquared.com/2013/05/why-data-driven-sales-forecasts-are-important/