Machine Learning Engineer Nanodegree

Capstone Project Proposal

audi-employee195
April 2019

Domain Background

Sales predictions are a very long research field in descriptive statistics and also lately by using machine learning approaches as well. This project will be about a challenging time-series dataset consisting of daily sales data of products and shops by the russian software firm 1C Company. This dataset is provided for a kaggle competition as well (see: https://www.kaggle.com/c/competitive-data-science-predict-future-sales/overview). Thus, the domain background is set and the information as well as the data are provided there.

Since it was not possible to use a company internal confidential data set applying a similar approach, we conducted research on publicly available data sets for demand/sales predictions. The structure and amount of data are comparable and thus will serve very well for this approach.

Problem Statement

In order to know which products C1 company needs to have in stock in which store, it is highly relevant to forecast the sales numbers for every product of the companies portfolio. The better the prediction the lower costs like running the storage, transportation costs for potentially unnecessary larger batch sizes of certain products, etc. Here an appropriate and well adjusted prediction algorithm would help the company. One can apply many time series regressions to predict the sales forecasts for the next month. The competition requires a prediction for the following month after the data set ends.

Datasets and Inputs

The dataset can be found here https://www.kaggle.com/c/competitive-data-science-predict-future-sales/data . The relevant files are

- item_categories.csv
- items.csv
- shops.csv
- sales_train_v2.csv
- test.csv

The size of all files (excluding sample_submission) is 99.4MB.
To specify the columns of the files see the following table:

File	Columns	Description
item_categories.csv	item_category_name, item_category_id	name of item category, unique identifier of item category
Items.csv	item_name, item_id, item_category_id	name of item, unique identifier of a product, unique identifier of item category
Shops.csv	shop_name, shop_id	name of shop, unique identifier of a shop,
sales_train_v2.csv	Date, date_block_num, shop_id, item_id, item_price, item_cnt_day	date in format dd/mm/yyyy, a consecutive month number, used for convenience. January 2013 is 0, February 2013 is 1,..., October 2015 is 33, unique identifier of a shop, unique identifier of a product, current price of an item, number of products sold.
Test.csv	ID, shop_id, item_id	an Id that represents a (Shop, Item) tuple within the test set, unique identifier of a shop, unique identifier of a product

The file „sales_train_v2.csv“ contains the most relevant information for the predictions. It shows the historical sales data from January 2013 until October 2015 per day per item per shop and its price. That means we have a sales history of 2 year and 10 months to learn from, in order to predict the amount of sales per item for the next month (November 2015).
The list of shops and products slightly changes every month.

Solution Statement

The approach towards the solution of the sales prediction problem will be conducted by applying various time series methods. The goal is to apply regression models that have been taught in the course and in addition to that further models.
The list of algorithms is:

- Linear regression as benchmark model
- XGBoostRegressor
- AdaboostRegressor (ensemble)
- LSTM (Long short term Memory)

Creating a robust model that can handle slight monthly changes of products and shops is part of the challenge. This will lead to detailed analysis of the data and might not only make regression a possible approach for prediction. A classification if or of not a product will be sold in November 2015 and if or if not a shop will sell are interesting issues that need investigation.

Benchmark Model

The benchmark model will be separated in two parts. The first benckmark model will be a linear regression of the monthly product sales and the prediction for the desired month November 2015. The second part will be the result of the submission in the kaggle challenge where the score is measured using the RMSE (rooted mean squared error).

Currently the median of the RMSE on 35% of the test data set (comparable to validation data set) is 1.09. The goal is definitely to be better than this median RMSE.

Evaluation Metrics

We will use the RMSE for the evaluation of the different predictions with the benchmark model but also with the submissions on kaggle. This is a quantifiable measure that can be used very well to compare predictions.
An initial evaluation of the benchmark model (linear regression) will be done in the kaggle competition to identify the RMSE of our benchmark model. As a next step the RMSE will be used to evaluate the previously mentioned algorithms against this benchmark RMSE.

Project Design

The following steps will be taken to achieve the goal of this project:

1. Data preparation
2. Data preprocessing for the various models
3. Training on different algorithms
4. Predictions on test data
5. Evaluation and comparison of the algorithms

Step 1 contains sub tasks like loading the data into a jupyter notebook and understanding the data for further analysis. Different visualizations on the data will be conducted to create a better understanding of the data. Especially the already mentioned case of varying shop and product numbers per month need to be examined carefully. In addition various standard statistical measures (e.g. mean, average, quartiles, trends) on the “sales_train_v2.csv” data need to be applied to understand the data in detail.
The second step will include sub steps like handling NaN values or One-Hot-Encoding for categorical variables (if necessary), splitting the training data into train and test set. If the various models need the data to be prepared in different formats these preparations will be done and described in detail in the report. Especially for LSTMs the data need to be prepared in periodic steps and aggregated for good results. In addition an aggregation of the training data per month will also be implemented since the goal is to predict sales numbers for a complete month.
Training on the training subset of the “sales_train_v2.csv” file will be the next step for all algorithms. The goal is to train all algorithms on the same training subset and test on the same test set (not to be confused with the “test.csv” file from the original data set). The fitting and training step of the models should be conducted at the same time and prepare for the next step,
Which will be the main task – the predictions on the test data, which have not been included in the training phase. The predictions then need to be adjusted depending on the level of aggregation to show the final result of item_id – predicted_sales_amount pairs for November 2015. Prediction on different data aggregation levels will be conducted (daily bases, weekly bases, monthly bases) and evaluated based on the result.
The Evaluation and certain adjustments to the hyper parameters of the algorithms will be conducted in the evaluation phase. It is important to start with the default setup of the algorithms and later on tune them based on their results. It can also happen that certain algorithms perform very poorly and thus will be excluded from the final optimization. Since there are just a few submissions per day possible to compare with the kaggle results, the “off-kaggle”-optimization and evaluation will happen first and the best model will be uploaded to the kaggle platform for comparison, which will also be stated in the capstone report.