Retail Data Analytics
Udacity Machine Learning Engineering Nanodegree
Capstone Proposal

Patrick Brus
Email: brus.patrick63@gmail.com

I. Domain Background

Retail Data Analytics (RDA) is used nowadays from shops in order to better predict the amount of articles, that might get sold and therefore to better estimate how much articles should be produced. This is very important, because the amount of sold articles can vary largely during the year. For example people are tend to buy more things before Christmas then during a normal, not holiday, week. This can be easily seen on the Amazon quarterly revenue on Statista [1]. The quarterly revenue of Amazon is always the largest for the fourth quarter, which indicates, that the people are consuming more during the fourth quarter than during the others. This is clear due to the fact that Christmas is within the fourth quarter and also the Black Friday, which leads to large worldwide consume, too. If a shop has too few products before Christmas, he would loose potential income. But if a shop has too much products, too much storage would be required and storage also costs money, so the company would again loose money. RDA can therefore be used in order to try to optimize the production of products, such that there is always an optimal amount available.

II. Problem Statement

The goal is to predict the department wide sales for each store for the following year. This should then help to optimize the manufacturing process and therefore to increase income while lowering costs. It should be possible to feed in past sales data from a department and to get the predicted sales for the following year.

III. Dataset and Inputs

The RDA dataset from Kaggle [2] is used for this project. The dataset contains historical sales data from 45 stores located in different regions. Each store is further divided into departments. The data itself is stored in an excel sheet. The excel sheet contains three tabs. The first tab contains the data from the stores. The second contains the features and the third contains the sales data.

A. Stores

There is data of 45 stores in total. Every store has it's own type and size, which is also included in the excel sheet. The information contained in the excel sheet is anonymized. Table I contains the statistics of the size from the stores. In total there are three different types of stores (A, B, and C).

	Size of Store
mean	130287.6
std	63825.3
min	34875.0
max	219622.0

Table I
Statistics of Store Size

B. Features

The features are related to a store. Table II contains all available features and a short description of each one, while table III contains the statistics to some of the features. The data for MarkDown1 - MarkDown5 is very incomplete and has to be dropped or different methods for handling missing data have to be applied.

Feature	Description
Store	the store number
Date	the week
Temperature	average temperature in the region
Fuel Price	cost of fuel in the region
MarkDown 1-5	anonymized data related to promotional markdowns
CPI	the consumer price index
Unemployment	the unemployment rate
IsHoliday	whether the week is a special holiday week or not

Table II
Features of the Dataset

	Temperature	Fuel Price	CPI	Unemployment
mean	59.4	3.4	172.5	7.8
std	18.7	0.4	39.7	1.9
min	-7.3	2.5	126.1	3.7
max	102	4.5	229	14.3

Table III
Statistics of Features

C. Sales

Each store also has historical sales data stored in the dataset. The sales data was collected from the fifth February 2010 until the first November 2012. Table IV contains all features related to the sales data and table V contains some statistics of the weekly sales.

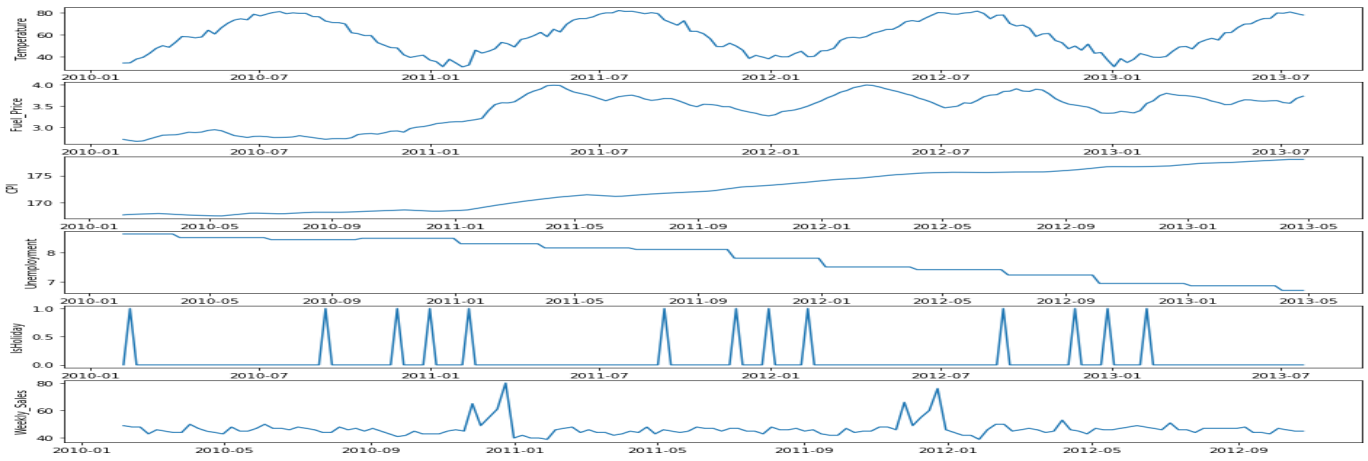

Figure 1. The historical data of features.

Feature	Description
Store	the store number
Dept	the department number
Date	the week
Weekly_Sales	sales for the department in the given store
IsHoliday	whether the week is a special holiday week or not

Table IV
FEATURES OF SALES DATA

	Weekly Sales
mean	15981.3
std	22711.2
min	-4989.0
max	693099.4

Table V
STATISTICS OF SALES

D. Discussion

This chapter takes a more detailed view on the data and all features. Figure 1 shows a time series analysis of some features. The first row contains the plots for the temperature. The temperature is alternating year by year having it's maximum somewhere around July and it's minimum somewhere in January. The temperature itself doesn't seem to have any influence on the weekly sales. The second row of the figure contains the fuel price. The fuel price had a large increase from January 2011 to July 2011. Afterwards, the fuel price is oscillating up and down. The third row of the figure contains the Consumer Price Index (CPI). This was steadily increasing since begin of the time series. In the next row one can see the Unemployment rate, which is steadily decreasing since the beginning of the time series. The functions of CPI and Unemployment rate make totally sense, because when the people have more jobs, they have more money to buy things and therefore the CPI goes up, because the more demand, the larger the prices. The fifth row of the figure contains the Boolean data of whether it's a holiday week or not. The last row contains the weekly sales data. As one can see, the weekly sales data is not automatically larger if the week is a holiday week. The peaks of weekly sales are in November and December. The peak in November is possibly due to Black Friday, while the peak in December is possibly due to Christmas. In January the weekly sales are the lowest.

IV. Solution Statement

In order to solve the Problem stated in section II, machine learning shall be applied. Different state of the art machine learning techniques should be applied and the best performing should be used for the final application. Within the final application, an user should be able to enter a week and a store of interest and to get the predicted weekly sales as output. As Machine Learning algorithms, the following ones shall be applied and evaluated:

- Linear Regression
- Decision Tree Regressor
- Random Forest Regressor
- XGBoost Regressor

The final model should be able to follow the pattern of weekly sales. It should be able to detect the peaks around Black Friday and Christmas and the low values in January.

V. Benchmark Model

As stated in the section IV, different state of the art machine learning algorithms should be applied and their performance should be compared. As benchmark model, the linear regression model should be used, because this one is very easy to interpret and to understand and easier models should be preferred over more complex ones. The linear regression library from Scikit-Learn shall be used in order to train the benchmark model.

Evaluation Metrics

The target data is numerical data. Therefore, root mean squared error (RMSE) can be used in order to get the best

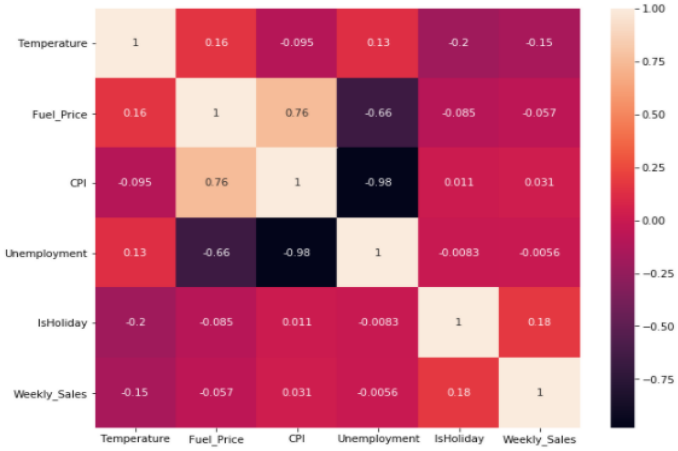

Figure 2. Heat map of Features and Weekly Sales.

performing machine learning technique. Equation 1 shows the formula for computing the RMSE.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{T}(y_i - \hat{y}_i)^2}{T}} \tag{1}$$

Project Outline

As first part, the data shall be visualized and analyzed in order to get a better understanding of all features and the data in general. The three separate data frames should be merged to one data frame using the store number as merging key. Then the correlation of all features to the weekly sales should be measured in order to only use relevant features and to remove too highly correlated ones. Figure 2 shows the correlation heat map from features to weekly sales. The correlation coefficient from store size to weekly sales is 0.85. As one can see, some important and possible features could be store size, Temperature and whether it's holiday or not. Additionally, a proper method for dealing with missing data should be applied in order to handle the large amount of missing data for the MarkDown1 to MarkDown5 features. When the data is ready, it should get split into training and testing data. The testing data itself is not used for training and is only used in the end to check the performance of the trained model on data it hasn't seen before. The training set should then be split into training data and validation data during training, using k-fold cross validation. K-fold cross validation helps to reduce the risk of overfitting and overly optimistic results. The training should be applied for all machine learning techniques stated in section IV. The Scikit-Learn library and the Amazon Sagemaker Library shall be used in order to implement and train all the models. The hyperparameters of the models should be optimized using the Bayesian optimization approach.

References

[1] Statista, "Net revenue of amazon from 1st quarter 2007 to 3rd quarter 2020," https://www.statista.com/statistics/273963/quarterly-revenue-of-amazoncom/, accessed 21-12-2020.

[2] Kaggle, "Retail data analytics," https://www.kaggle.com/manjeetsingh/retaildataset, accessed 21-12-2020.