Data Mining 2

Data Mining CUP

Presented by
Group 4
Adila Aghazada
Abdullah Al Murad Chowdhury
Bengi Koseoglu
Khizer Naushad
Na Gong
Qian Xia
Sanjita Suresh

Submitted to the
Data and Web Science Group
Prof. Dr. Heiko Paulheim
University of Mannheim

May 2018

1 Introduction

This year's Data Mining Cup (DMC 2018) was about predicting the sold out date of products based on the stock at the beginning of the month and sales unit of each day. For the cup, a real world data from a sports company had been provided. The problem was a time series task based on sales unit in each day, but additional information such as color, brand, size and price of the product had also been provided. In order to evaluate the sold out units and sold out date, the sales data from October to December(3 months) was used for training and tested on the January data. This report serves as a documentation of our team's approach to solve the task. For the project Python, R and RapidMiner had been used.

2 Data Description

The DMC 2018 provides three individual datasets (items, prices and train) in .csv format. The 'items' dataset serves as master data, contains all necessary information to describe an item. In addition, it also includes opening stock per distinct product for the month of February, 2018. The 'prices' dataset provides historical item's price information between October, 2017 and February, 2018. The final dataset 'train' contains daily sales transaction from October 2017 to January 2018. Preliminary data analysis has been made before starting modeling in order to get to know the dataset better and to get a deeper understanding of the problem at hand.
The color distribution has been analyzed first and it has been found out that the majority of the items are of colors schwarz (black), followed by blau (blue), weiss (white) and rot (red).
Secondly, the brand distribution of the products have been analyzed by plotting pie chart. As it can be seen that 'Nike' mostly dominates the entire sales transactions (55%) followed by 'adidas' (35.1%). In contrast 'Under Armour' contributes a minor portion (0.4%) in daily sales.
Then, sales trend have been analyzed by plotting daily average sale on y axis with the dates on x axis. It is clearly be observed that the sales volume has skyrocketed and the highest numbers of items sold on Black Friday. Another peek was shown during the Shopping Day in Germany. The sales unit remained almost same on other business days with some fluctuations.

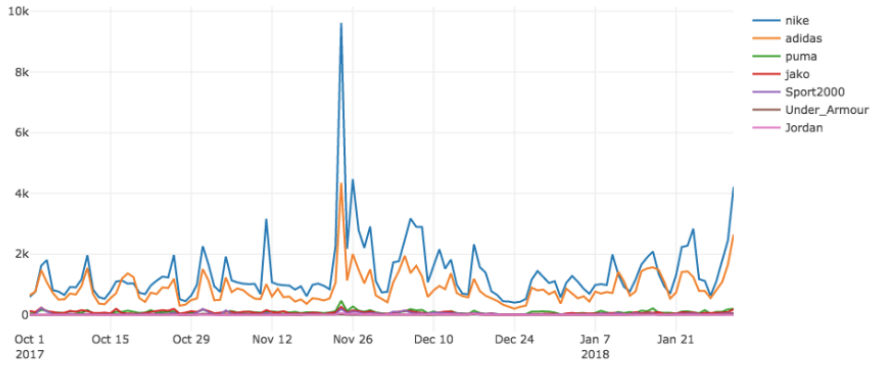

Figure 1: Sales trend

3 Preprocessing steps

3.1 Data Cleaning

By analyzing the dataset, the *price* attributes of products per day had missing values in about 153 columns. The approach of calculating the average (mean) values based on each product was used to replace all the missing values in rows. About the missing values in *Subcategory*, a new subcategory was created and all the missed products were assigned to the new subcategory. For the *size*, the most frequent size of each brand was found to fill the missing value. Furthermore, the original dataset is highly structured and cleaned but just two abnormal values were detected in sales units and several abnormal peaks in products recommended prices. By considering their tiny proportion and the intangible effect of abnormal prices, all abnormal values were kept as original.

3.2 Feature Engineering

In order to increase the effectiveness of the later predictive models, 18 new features with about 200 dummy features were created as following based on the deep domain knowledge gained from given dataset and the implementation of regression models:
new_size: Unified unstructured size format of all 26 brands with consideration of different products types and customer category.
avg/mediann_temp: Gain the average daily temperature (Oct-Feb) of 8

German cities where the '11teamsport' company is located from the weather website[1].
company_offer: A binary variable to show whether the day has the special discount offered by 11teamsport company. The company that provided the dataset was discovered by team 5.
day/month/year/ holiday/weekday: Flatten the sales date to three individual variables and added a binary holiday with seven weekday features because most holidays and Sunday is non-business day in Germany.
new_product: In terms of the boosting impact of new products on sales, a binary product life cycle feature was created based on the released date of each products.
price_daily_change: Due to the short reflection time of sales on price in fashion industry, a percentage variable was computed to show the daily price change of each product for the previous day.
sum_units: The sum of previous month sales for each products.
dummy features: In order to efficiently fit all prediction model especially regression models, all categorical variables (new_size, color, brand, category, mainCategory, subCategory) were converted to dummy features.

3.3 Feature Selection

Having considered the large number of attributes in our dataset which may hinder the efficiency of the algorithm, feature selection and dimension reduction played an important role before the modeling process. Therefore, four selection methods were used to select attributes:

- Mutual Information
- Recursive Feature Elimination
- F-Classification
- PCA-Principal Component Analysis

First the top 20 attributes from each selection methods were picked out. The final results were tabulated based on the number of times the individual variable was selected in order to generate an initial feature relevancy ranking (a sample is shown in Table1). At the end, 34 variables were selected.

[1] www.accuweather.com

Selected by Three Algorithms	Selected by Two Algorithms
mainCategory15	mainCategory1
mainCategory9	Category10
Category16	Category2
Category18	Category33
Category7	Category37
subCategory0	subCategory16
subCategory32	subCategory3
	newproduct

Table 1: Selected Features by Feature Selection

4 Modeling

In the time series, it's assumed that preceding values will have an affect on the current value. One way to account for this is to apply models that are capable of capturing different temporal structures and seasonality such Autoregressive Integrated Moving Average Model(ARIMA). Other way is to treat time series forecast as a regression problem while taking preceding sales into account such as windowing. Third way is to predict daily unit sales for each day by using regression and machine learning methodologies. As suggested by Prof. Dr. Paulheim, the baseline has been chosen as 15th of January, and the error rate was ca lculated as 261.02.

4.1 Clustering

Instead of predicting the sales for whole products or for individual products, products can be grouped according to their similarities and models can be built on top clusters. Based on different criterions, 5 clustering approaches were used before modeling: Brand Binning, Category Binning, Sales Binning, K-means Clustering and Time Series Clustering.
Brand clustering was based on the number of products in each brand, and whole dataset was divided into 5 clusters: Nike, Adidas, PUMA, Jako and other brands. Category clustering was divided into 3 based on the main category 1 9 and 15. In sales clustering, the dataset was divided into low, average, good, top and new sales products in January. Before K-means clustering, calinski_harabaz_score and Elbow analysis were used to decide the optimal number of clusters. The results were 7 and 12 respectively. Having tried both, K-means clustering with 7 clusters was selected.

4.1.1 Time Series Clustering

The above mentioned clusters don't take sales variations in time into account, therefore a new clustering methodology specific to time series applications have been used by taking only sales of each product between October and December as input. This clustering methodology includes using Dynamic Time Warping(DTW), which is a similarity measure that allows comparisons of two time-series sequences, coupled with hierarchical clustering [1]. DTW was calculated using R, and the resulted distance matrix passed to the hiearchical clustering. The optimum number of clusters have been decided according to the elbow and calinski harabasz measures, and the tree was cut at 7, to generate 7 clusters.

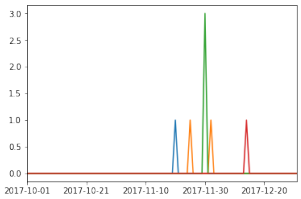

Figure 2: Cluster1

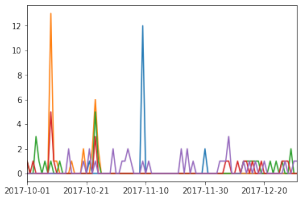

Figure 3: Cluster2

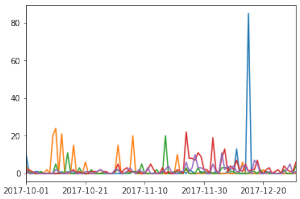

Figure 4: Cluster3

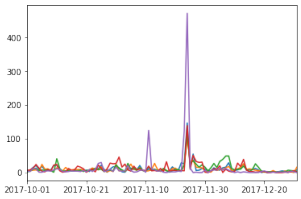

Figure 5: Cluster4

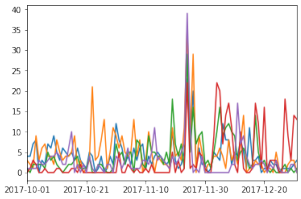

Figure 6: Cluster5

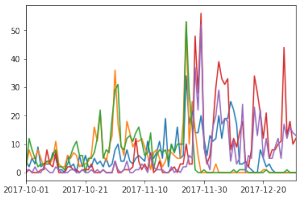

Figure 7: Cluster6

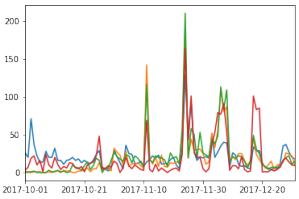

Figure 8: Cluster7

Above you may found the generated cluster's sales trend over time. In order to represent their trend clearly, samples are chosen from each cluster.

4.1.2 Clustering Comparison

In order to choose the best clustering for our dataset, GBoost Regression without parameter tuning have been applied by assigning 31st of January as fall out date and tested with test0. As shown in Table2, K-means clustering resulted the worst whereas brand and category bining performed almost the same and time series clustering resulted as the best, with an error rate of 261 which is just around our baseline. With time series clustering, the error rate instantly dropped 15 day, therefore we decided to use time series clustering in our models.

Clustering	**Performace**
K-means	280
mainCategory	276.6
Brand	276.1
Time Series	261

Table 2: Performance of different clustering approaches.

4.2 ARIMA

ARIMA is a time series model that captures suite of different standard temporal structures in time series data. The model has been applied both to the whole dataset and to the clustered dataset in RapidMiner using the operators moving average, arima trainer and apply forecast. The training data was sent to the moving average operator with the window width of 5 and average as aggregation function based on the 'units' feature. Then ARIMA trainer modeled the data with 'units' as the time series attribute with the default parameters of p=1, d=0 and q=1. The units for next 31 days of January were predicted using the apply forecast operator. The error from the ARIMA model was 262,034 and on the time series cluster was 248,928. ARIMA resulted with the lowest day difference among all our individual models.

4.3 Windowing Approach

Windowing is an approach that takes the preceding values into account while predicting the current value. The approach has been applied individually for each product and on top of time series clusters by taking the mean daily sales in the cluster and predicting the future mean sales. On individual level, linear regression with window size of 1 and 30 have been applied by focusing

on only sales. With 30, Artificial Neural Network (ANN) also applied to compare the success of different algorithms. Additionally, weather data has been included into the model with window size of 1, by using regression and ANN in order to see the effect of weather on sales. As suspected, the error rate decreased (253) which is an indication of correlation between weather and sales. On clustering level, time series clustering has been used with window sizes of 1, 5, 10, 15, 20 and 30. The error rate of 252.6 have been observed with window size 20 and 252.8 with 30. Therefore, ANN applied with window size of 20 and ANN, gradiant boost, random forest applied with window size of 30. With the combination of window size 20 and ANN, the lowest error rate (250) has been reached.

4.4 Regression Model

Gradient tree boosting is a robust technique that stands out in many machine learning applications [2]. This study adopted two tree boosting-based learning systems Gradient Boosting Regression (GBR) and XGBoost Regression (XGBR) whose strong learning liability is widely recognized in Kaggle competitions [2]. During construction, the regression models were first integrated with time clusters by building separated models for each cluster. Then, different parameters were tuned with GridSearch algorithm for almost every cluster. Due to the run time barrier, 'two clusters' parameters were optimized manually. Based on results of feature selection, total 12 regression models were trained with different features. Both regression models performed quite similar with a slight competition of GBR. The top three performances are shown in next section.

4.5 Combined Model

Model	C1	C2	C3	C4	C5	C6	C7	C8
ARIMA	191.93	110.69	42.80	7.82	22.26	14.74	3.72	101.29
Wondowing	192.23	114.71	41.81	6.32	22.88	15.03	4.0	101.29
GBR	200.14	107.5	41.21	6.63	21.66	14.98	3.57	101.29
GBR_sum	199.13	105.12	41.21	7.38	21.33	12.65	3.74	101.29
XGBR	201.41	107.19	40.66	6.96	21.22	14.58	3.70	101.29

Table 3: Performance of individual models per cluster.

After all above five individual models, the best models of each cluster were combined together to integrate the learning advantages from every model. According to the individual performance of clusters in Table 3, four model

combinations were tested. As a result, the configuration of the best combined model is: cluster 1. ARIMA, cluster 2. GBR w/ sumUnits, cluster 3. XGBR cluster, 4.Windowing w/ ANN (lag20), cluster 5. XGBR, cluster 6. GBR w/ sumUnits, cluster 7. Windowing w/ LR (lag30), cluster 8. GBR.

4.6 Ensembling Model

After finishing all individual models, stack ensembling was tried to improve the result. The attributes were the prediction results from ARIMA, regression models, Windowing approach and combined models. XGboost regression was used to predict the final sales. The stack ensembling based on 7 clusters and whole dataset were tried with different results combination and parameters. But the best result was about 250.85, which was higher than the individual models.

5 Results and Conclusion

5.1 Result Analysis

Based on the error function provided by DMC2018 cup and standard deviation metric, a statistical performance comparison of different models is visualized in Table 4.

Model	Macro Avg. Score	Standard Deviation
ARIMA	248.9	0,34
Windowing-lag20 w/ ANN	250.8	0,38
GBoost Regression	253.6	0,59
GBR w/ sumUnits	251.7	0,45
XGBoost Regression	254.4	0,66
Combine Model	245.8	0,37
Ensemble Model	250.9	0,61

Table 4: Performance evaluation of different models.

According the analysis, following results can be induced:
i. All seven models successfully beat the baseline of 261.02 by 10 days on average.
ii. Three regression models have quite similar learning capability but the GBR with the feature of *sum_units_of_previous_month* has a slight more competition by about 2 days deviation. This also indicates that the sale of previous month is indeed a good index to forecast the sales of the coming

month in sporting goods retailer.
iii. The windowing approach which applying both Linear Regression and ANN has better performance than pure regressors. Moreover the time-series model ARIMA perform the best among all the individual models. This is additional evidence which approve the shining learning competition of ARIMA in time series analysis.
iv. In general, by avoiding the weakness of each individual model and integrating their strengths only, the combined model finally stands out from all models with the lowest prediction error (245.8).
v. Ensemble model built from XGBR gives an acceptable result but not robust as expected. This could be affected by the limited parameter tuning.

5.2 Conclusion

In this study, the sold out date of each products was successfully predicted by applying Windowing, ARIMA, Regression, Ensemble and Combined models on historical sales data issued by DMC2018 competition. In terms of a comprehensive performance evaluation, Combined Model is regarded as the best model in this prediction of sold out date. In current project, the final ensemble model is built on XGBR model with whole dataset and limited by the inefficient parameter tuning due to the shortage of hardware and time frame. It is could be improved by further model and parameter optimization.

Bibliography

[1] Luk, M. (2017). Dynamic Time Warping: Time Series Analysis II. Retrieved May 19, 2018, from `https://sflscientific.com/data-science-blog/2016/6/3/dynamic-time-warping-time-series-analysis-ii`

[2] Chen, T.Q., Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Retrieved May 19, 2018, from `http://delivery.acm.org/10.1145/2940000/2939785/p785-chen.pdf?ip=134.155.219.159&id=2939785&acc=CHORUS&key=2BA2C432AB83DA15`

Appendix

Name	Matrikel Nr.	Data Exploration	Data Preprocessing	Clustering Model	Prediciton Model	Presentation	Report
Adila Aghazada	1622160	★★★	★★★	☆☆☆	☆☆☆	★★★	★★★
Murad Chowdhury	1623027	★★★	★★★	★★★	★☆☆	★★★	★★★
Bengi Koseoglu	1619463	★★★	★★★	★★★	★★★	★★★	★★★
Khizer Naushad	1622808	★★★	★★★	☆☆☆	★★☆	★★★	★★★
Na Gong	1616497	★★★	★★★	★★★	★★★	★★★	★★★
Qian Xia	1619383	★★★	★★★	★★★	★★★	★★★	★★★
Sanjita Suresh	1583940	★★★	★★★	☆☆☆	★★★	★★★	★★★

Figure 9: Contributions of team members