Data Science Techniques and Applications
Coursework I Dataset analysis

Jose Manuel Magaña Arias

March 11, 2019

MSc Data Science

Birkbeck College, University of London

Abstract

In this short paper I present the first part of the coursework from the Data Science Techniques and Application module (DSTA). The goal of this first stage is to build a basic analysis of a Kaggle dataset. In Phase I, I present the different Kaggle repositories that were considered to select a dataset. I also explain the rational behind my decision to selecte the *'Walmart Recruiting - Store Sales Forecasting'* dataset. In Phase II, I provide a brief description of the contents of the datasete and information about the Kaggle challenge proposed for this dataset. Finally in Phase III, I perform a dimensional analysis of the features and discuss the range, quality and possible issues of the data. This analysis serves as a preparation for coursework II where dimensionality reduction techniques such as principal component analysis (PCA) will be applied.

Phase I - Subscribe to Kaggle and select a dataset

Kaggle is an online data science community that allows users to find and publish datasets, perform data analytics and build, test and share models[3]. One of the most popular Kaggle's services are the machine learning public competitions where companies and research institutes post real-world problems so data scientist have the opportunity to compete to build the best algorithm. In order to be part of this community one has to sign in and create a profile. I have done this by using my academic email address `@dcs.bbk.ac.uk`. My profile can be accesed via the following link: `https://www.kaggle.com/jmagan01`

In order to select a dataset for analysis two main repositories were considered:

1. The Kaggle's public dataset platform which comprises a collection of almost 15,000 datasets freely available to the public. Not all of those datasets are necessarily part of a specific Kaggle competition. Kaggle's staff suggests to explore the short list of 'Machine learning friendly' datasets[1] which was created with the aim to encourage new data scientist to start exploring datasets.

[1] https://www.kaggle.com/annavictoria/ml-friendly-public-datasets?utm_medium=email&utm_source=intercom&utm_campaign=data+projects+onboarding

2. Datasets from active and expired Kaggle competitions were revised. [2]

After a careful review, the **Walmart Recruiting - Store Sales Forecasting**[3] was selected. The main reasons behind this decison are:

1. This dataset was part of a serious Kaggle competition in the past, which means it will provide interesting challenges in the absence of ideal data i.e. missing values, outliers, etc.
2. It includes both, internal information (e.g. historical sales from Waltmart stores and discount activities) and external features such as consumer price index, unemployment rates and fuel prices.
3. This is a time-series dataset, which provides a good suit for my strong interest in the topic of *time-series forecasting* which is also the area of my MSc graduation project where I propose to benchmark traditional time-series forecasting techniques such as AutoRegressive Integrated Moving Average (arima) models against modern methods like Recurrent Neural Networks (RNNs).

Phase II - The 'Walmart Recruiting - Store sales forecasting' dataset

This dataset is part of the (expired) Kaggle competition *Walmart Recruiting - Store Sales Forecasting* from 2014. This was not only a machine learning competition but a recruiting strategy from Walmart. This dataset provides historical sales data from 45 Walmart stores located in a number of regions of the United States. Each store has a number of departments, the ask of the competition is to forecast (predict) the sales for each store and department by week.

Walmart runs some discount events (markdown events) throughout the year, which usually preced prominent holidays such as Super Bowl or Christmas. These markdowns are known to have an impact on sales (markdown effect) but it is rather difficult to predict which departments are affected and also to quantify the impact in advance. The markdown information is provided as anonymized data related to promotional markdowns during those particular weeks. Other interesting features related to the region of the store for the given dates are also provided, i.e. average temperature in the region, the unemployment rate, the cost of fuel in the region and the consumer price index (CPI).

The dataset is comprised of four files:

1. **train.csv**: training data (historical sales) broken down by week, store and department. It includes a flag indicating if the week is a holiday week or not.
2. **test.csv:** same as train.csv except for that the weekly sales have been cleared out. This is the data to be forecasted.
3. **stores.csv:** anonymized data from 45 stores, including the type and size of the store.
4. **features.csv:** anonymized markdown activity data by store. It also includes additional data related to the region of the store (temperature, fuel price, CPI and unemployment rate).

[2] https://www.kaggle.com/competitions
[3] https://www.kaggle.com/c/walmart-recruiting-store-sales-forecasting

Phase III - Dimensional Analysis

This dataset has one forecast variable (also called dependent variable) which is the future sales value. There are two variables of control: store and department, while the rest of the features or dimentions can be considered regressors (also known as independent variables or explicative variables) which might have an impact on future sales value. Some of the definitions below have been taken directly from the Kaggle website.[4]

Weekly_Sales - this variable captures the historical sales value for the given week, department and store in US$. The training dataset is made of 8,190 data points, there are 45 different stores meaning there are about 182 data points per store. The average weekly sale value is close to $16,000 with range $[-4988, 693099]$. In order to deal (remove) any negative sales values a data cleansing routine will be needed before conducting any data analytics.

Store - the store number ranging from 1 to 45.

Dept - the department number ranging from 1 to 99.

Date - time variable to identify the week. It covers the period from 05-Feb-2010 to 01-Nov-2012. It is possible to find gaps in the data i.e. missing weeks. Pratama et al. present a review of missing values handling methods on time-series data which might be considered.[7]

Temperature - average temperature in the region (in Fahrenheit degrees[2]). This dimention has a range of $[-7.29, 102]$with a mean of 59.4 °F. The quality of the data is good, there are no missing values neither extreme outliers. Temperature might have an impact on sales on certain products and departments especially during very warm or cold seasons of the year.

Fuel_Price - this dimention captures the cost of fuel in the region, it has a range of $[2.47, 4.47]$with a mean of 3.4 US$. The quality of the data is good, there are no missing values neither extreme outliers.

MarkDown1-5 - There are 5 different MarkDown dimensions, this is anonymized data related to promotional markdowns that Waltmart has applied. This data is presentde by week and store, so there is no detail at the department level, it is only available after November 2011, and is not available for all stores all the time. Missing values are marked as *NA*. There are some negative values which make no sense from the business point of view, so a data cleansing routine will be needed in order to deal with those data points. As an example, the first MarkDown dimention has a range $[-2783, 10385]$ with mean equal to 7,032 US$.

CPI - the consumer price index is defined by the US Bureau Labor of Statistics as follows:

"The Consumer Price Index (CPI) ia measure of the average change over time in the prices paid by urban consumers for a market basket of consumer goods and services."[5]

This data is presented in a weekly frequency for each region where each of the 45 Walmart stores are located. It rages$[126.0, 228.9]$ with a mean equal to 172.5. An inverse correlation between CPI and sales values is expected.

Unemployment - the unemployment rate is defined by the US Bureau Labor of Statistics as follows:

"The national unemployment rate reflects the number of unemployed people as a percentage of the labor force. People are classified as unemployed if they do not have a job, have actively looked for work in the prior 4 weeks, and are currently available for work."[6]

[4]https://www.kaggle.com/c/walmart-recruiting-store-sales-forecasting

The unemployment data is updated every quarter, so the figures are repeted for those weeks within the same quarter. This indicator is presented for each region where each of the 45 Walmart stores are located. It rages $[3.68.0, 14.3]$ with a mean equal to 7.8. The expectation is to find an inverse relationship between unemployment rates and sales growth.

For both variables, the CPI and Unemployment rate, the data is missing for 2013 and beyond so in order to apply a predictive model for this period of time the data will need to be updated.

IsHoliday - this is a boolean variable to identify if the week is a relevant holiday week or not. In the training dataset the following dates indicate a week with relevant holidays.

Event	2010	2011	2012
Super Bowl	**12-Feb**	**11-Feb**	**10-Feb**
Labor Day	**10-Sep**	**09-Sep**	**07-Sep**
Thanksgiving	**26-Nov**	**25-Nov**	**Nil**
Christmas	**31-Dec**	**30-Dec**	**Nil**

Historical sales values as features - In signal and time series analysis the historical (lagged) values of the variable of interests (weekly sales) are usually correlated with future sales values, this is called *Autocorrelation*[1] or *serial correlation*. George Box and Gwilym Jenkins proposed a methodology to model an event as a function of its past values with the assumption that the patterns (i.e. trend, seasonality and cycle) will persist in the future. In order to apply those techniques the lagged values should be treated as additional features (or dimensions) to explain future sales values. This data is not explicity provided in the dataset, but can be easily computed from the historical data. The Box-Jenkins methodology was published in 1970 in a famous book called *Time Series Analysis, Forecasting and Control*[4].

References

[1] Wikipedia contributors. Autocorrelation. in wikipedia, the free encyclopedia. `https://en.wikipedia.org/wiki/Autocorrelation`, 2019. Last accessed 10 March 2019.

[2] Wikipedia contributors. Fahrenheit. in wikipedia, the free encyclopedia. `https://en.wikipedia.org/wiki/Fahrenheit`, 2019. Last accessed 10 March 2019.

[3] Wikipedia contributors. Kaggle. in wikipedia, the free encyclopedia. `https://en.wikipedia.org/w/index.php?title=Kaggle`, 2019. Last accessed 10 March 2019.

[4] Gwilym M. Jenkins George E.P. Box. *Time Series Analysis, Forecasting and Control*. Holden-Day, San Francisco, CA, 1st edition, 1970.

[5] US Department of Labor. Consumer price index. in bureau of labor statistics. `https://www.bls.gov/cpi/`, 2019. Last accessed 10 March 2019.

[6] US Department of Labor. Labor force statistics from the current population survey. how the government measures unemployment. `https://www.bls.gov/cps/cps_htgm.htm`, 2019. Last accessed 10 March 2019.

[7] Irfan Pratama, Adhistya Permanasari, Igi Ardiyanto, and Rini Indrayani. A review of missing values handling methods on time-series data. pages 1–6, 10 2016.