DATA70002 Understanding Data and its Environment

Coursework Project

This coursework project is mainly concerned with **data pre-processing** for a sales forecasting problem across different stores in the retail industry. The task involves the preparation and analysis of historical sales data collected from a large drug store chain in Europe (like Boots in the UK) for forecasting. The aim is to expose you to a realistic business case and to gain understanding and insight about some of the ways in which data can be fully prepared to optimise its analytical value.

Assessment and submissions

- Deadline for individual report (55% of the marks for the module) submission: 3.00pm 15th June 2020

Please find further requirements and tips in the following pages.

Description of the business context

Accurately forecasting sales is one of the most difficult challenges faced by retailers worldwide, as sales are influenced by many factors, such as promotions, competition, holidays, seasonality and locality. In this project, the overall business objective is to predict 6 weeks of daily sales for 1,115 drug stores located across Germany, as reliable sales forecasts enable store managers to increase the overall productivity and profitability of the retail business and improve their customer satisfaction.

However, the challenges in this sales forecasting problem are to take into account various types of factors and to deal with missing data from historical records. Thus, you are commissioned to get the historical dataset fully pre-processed for reliable and accurate forecasting, and the major data preparation tasks, like integration, description, visualisation, cleaning and transformation, should be explored. The available datasets are briefly introduced below. Note that some stores in the datasets were temporarily closed for refurbishment.

- **stores.csv**

This excel file contains the supplementary information for the 1,115 drug stores.

Column	Description
Store	the anonymised store number

StoreType	4 different store models: a, b, c, d
Assortment	an assortment level: a = basic, b = extra, c = extended
CompetitionDistance	distance in meters to the nearest competitor store
CompetitionOpenSinceMonth	the approximate month of the time when the nearest competitor was opened
CompetitionOpenSinceYear	the approximate year of the time when the nearest competitor was opened
Promo2	a continuing and consecutive promotion, e.g., a coupon based mailing campaign, for some stores: 0 = store is not participating, 1 = store is participating
Promo2SinceWeek	the calendar week when the store started participating in Promo2
Promo2SinceYear	the year when the store started participating in Promo2
PromoInterval	the consecutive intervals in which Promo2 is re-started, naming the months the promotion is started anew. e.g., "Feb,May,Aug,Nov" means each round of the coupon based mailing campaign starts in February, May, August, November of any given year for that store, as the coupons, mostly for a discount on certain products are usually valid for three months, and a new round of mail needs to be sent to customers just before those coupons have expired

- **train.csv**

This file contains the historical sales data, which covers sales from 01/01/2013 to 31/07/2015. It includes the following fields:

Column	Description
Store	the anonymised store number
DayOfWeek	the day of the week: 1 = Monday, 2 = Tuesday, …
Date	the given date

Sales	the turnover on a given day
Customers	the number of customers on a given day
Open	an indicator for whether the store was open on that day: 0 = closed, 1 = open
Promo	indicates whether a store is running a store-specific promo on that day
StateHoliday	indicates a state holiday. Normally all stores, with few exceptions, are closed on state holidays. Note that all schools are closed on public holidays and weekends. a = public holiday, b = Easter holiday, c = Christmas, 0 = none
SchoolHoliday	indicates if the (Store, Date) was affected by the closure of public schools

- **test.csv**

This file is identical to train.csv, except that Sales and Customers are unknown for the period of 01/08/2015 to 17/09/2015.

Evaluation of forecasting accuracy

In this project, the following Root Mean Square Percentage Error (RMSPE) or other appropriate errors can be used to evaluate forecasting accuracy, if you wish to build a forecasting model on the pre-processed datasets.

$$\text{RMSPE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{y_i - \hat{y}_i}{y_i}\right)^2}$$

where *N* is the total number of data records for accuracy evaluation, y_i is the actual sales for the *i*th record, and $\hat{y}_i$ is the sales forecast for the *i*th record. Note that zero actual sales should be dealt with appropriately.

General requirements: You are expected to collaborate with your group members to understand the business problem and lay out the data pre-processing plan for formative feedback, before you complete an individual report (length: 2500 words) for assessment.

Your work should cover (but not be limited to) the following.

- Review the data and describe them in terms of available variables, quality, and relevance to the sales forecasting,

- Integrate relevant data sets together as appropriate,

- Pre-process the data as appropriate for further analytics, for example, encode categorical variables, create new variables, identify and deal with missing values or records as appropriate.

- Identify the key variables affecting sales, for example, check whether competition and promotions have an impact on sales, and how public holidays cause sales fluctuations.

- Build a forecasting model (which can be a simple regression model or any advanced machine learning model you are familiar with). The main focus here is to check the validity of the variables you've pre-processed and identified above, rather than to build a sophisticated forecasting model.

- Interpret key results, assumptions and limitations of your data pre-processing and analysis.

Some tips for writing the report

- Imagine that you are writing the report for someone to read not simply to pass the course!
- A report should include an introduction and a conclusion. Marks are available for these two sections.
- A good report is a narrative; not simply a reporting of what you did.
- Your goal is to communicate your findings not simply to churn out the analyses.
- The steps above are components that should be included in the analysis and reporting; how you include them is up to you. Reports that simply use the task descriptions above as headings will lose marks.
- Distinction level reports tend to go beyond the specification – adding extra ideas connections /analyses or ways of presenting the data that are not specified above. I view these favourably (as long as they are well done!) but they are not essential.
- Put some effort into the layout and presentation – these are easy marks.
- Exploratory analysis should be included in the main report where appropriate and where it adds to the narrative. Assumption test output can be included in the appendices as can any exploratory analysis which adds to the story you trying to tell but would clutter up the main body of text.
- Strike the right balance between too few and too many charts and tables. One-two per page (depending on size) is a good rule of thumb.
- You should, in the conclusions, report on the limitations of the data you have used or on what future studies of the same topic might need to look for.
- You should label/number figures and tables fully and appropriately. A general rule of thumb is that a figures and diagrams should be understandable on their own without having to refer to the main text. Figures should be referred to them in the main text by “Figure n” or “Table n” where n is the number of the table or figure in the sequence through the paper. Note that the words “Table” and “Figure” have a capital first letter (as “Table 1” is a pronoun).
- Any plagiarism from source/reference material or other group's work will be penalised and may result in a mark of zero (please refer to your programme handbook).
- You must submit your coursework report for this course to Blackboard by the deadline.

An indicative breakdown of marks is listed in the following table

Assessed report	%
Introduction	15
Methodology (major data pre-processing tasks)	35
Results (description, discussion, analysis, etc.)	25
Conclusion, implications and recommendation	15
Layout and presentation	10