Project C5: Store Sales - Time Series Forecasting

CRISP-DM Report

Project members:
Juhan Pauklin, Joonas Tiitson. Kristjan Radsin

Business understanding

Background:

Our organisation, a Ecuadorian retail company called Favorita, is seeking to enhance its sales forecasting capabilities to optimise inventory management and improve overall operational efficiency. The company operates multiple stores, and it faces challenges in accurately predicting future sales, leading to suboptimal stock levels, excess inventory costs, and occasional stockouts. The ultimate goal is to leverage data-driven approaches to forecast store sales more accurately, thereby reducing costs and enhancing customer satisfaction.

Business goals:

Improve Inventory Management - Develop a robust predictive model to accurately forecast store sales, allowing the company to optimise inventory levels and reduce carrying costs.

Business Success Criteria:

Reduce the procurement of excess inventory. Excess inventory should not exceed 7.5% of sold inventory.

Inventory of resources:

The list of resources available for the project consists of first and foremost the data given to by Kaggle. This data includes seven datasets: Holiday_events.csv, oil.csv, sample_submission.csv, stores.csv, test.csv, train.csv, transactions.csv.

The software consists of the Jupiter Notebook and python moduls such as numpy, scikit-learn, matplotlib and others.

Resource of personnel is made out of 3 data science students.

Requirements, Assumptions, and Constraints:

- Requirements: Access to historical sales data, computing resources for modelling, and team communication.
- Assumptions: Sales patterns are influenced by factors like seasonality, promotions, and economic conditions such as oil price.
- Constraints: Time constraint. Otherwise there aren't any other real constraints.

Risks and Contingencies:

Main risk is one or more of the team members falling ill and being unable to participate due to that. This would hamper manpower and significantly slow down progress.

There is no great solution to falling ill. The damages could be mitigated by notifying other team members early and trying to complete the assigned task.

Terminology:

The terminology of this project will use common sales terms. The standardisation of such elements ensures the mutual understanding between anyone who stumbles upon this project.

Costs and benefits:

The cost involves the time and resources needed for model development, data processing and team collaboration. The benefit of this might include the optimisation of location, price and time of products in stores. Which improves the overall sales or spending of the storechain.

Data-mining goals:

The goals of this project are:

- To use different machine learning techniques to forecast the number of sales of a product based on the specifications of its status.
- To identify the main factors of what influences the sale of products, such as seasonality, promotions and external events (oil prices).

Data-mining success criteria:

Achieve a Mean Absolute Percentage Error (MAPE) of less than 10% in the forecasting models.

Extract meaningful insights from the data such that it can be used for creating future marketing and sales strategies for the store chain.

Data understanding

Gathering data:

The data necessary to fulfil our data-mining goals would include information about stores, sales, products, and potential external factors such as holidays or ongoing promotions. We would need to know info about things such as what was sold in which stores and on what day, as well as when holidays are celebrated. Luckily, all of this data already exists, as it was provided by Kaggle. The provided datasets will be the basis for our data-mining project.

One of the requirements for the data is that it can’t be too old - needs to be relevant in the present.

Describing data:

All the data we have has been sourced from the associated Kaggle page, and is in the csv table format. These files include stores.csv, transactions.csv, holidays_events.csv, and oil.csv.

The stores.csv file contains metadata about stores, containing 54 cases and 5 fields: store_nbr, city, state, type, cluster. The cluster field groups together similar stores.

The transactions.csv file contains 83488 cases and the following fields: date, store_nbr, and transaction fields, where the last field contains the amount of transaction on the given date.

The holidays_events.csv file contains data about holidays and events, containing 350 cases and 6 fields: date, type, locale, locale_name, description, transferred. The transferred field refers to whether a holiday has been moved to a different date by the government.

The oil.csv contains daily oil prices, 1218 cases and the following fields: date, dcoilwtico. The latter field is the price of oil at the given date.

There are also the train.csv and test.cvs files, which contain id, date, store_nbr, family, sales (only in train.csv), and onpromotion fields. The family field describes the type of product sold, the sales field the total sales for a given product family and the onpromotion field whether there was an ongoing promotion for that product family.

This data is sufficient for our goals, as it includes the information we are looking for as well as enough cases to build models upon.

Exploring data:

Looking at the data more closely, these are some of the initial observations we have been able to make. Most stores in our data are located in the cities Quito and Guayaquil, located in the states of Pichincha and Guayas. The stores are not grouped evenly in our dataset. Daily transactions at one store range from as low as 5 to as high as 8359. Daily oil prices range from 26.19 to 110.62. There were a few missing fields in the oil.csv file. The starting date for the recorded transactions and oil prices is 2013-01-01, while the end date is 2017-08-15 for transactions and 2017-08-31 for oil. Holidays and events start from 2012-03-02 and end at 2017-12-26. Given these dates, the range where we can perform model training is from 2013-01-01 to 2017-08-15.

Verifying data quality:

After examination, the data we have appears to be high-quality. We have data about the things we need to continue with our plan. The data has some missing values in the oil prices, however as this file contains a lot of cases and the missing values are spread out, they do not affect the overall quality of the data and if necessary, can be substituted. The rest of the data seems to be clean, and overall it contains the data we need to work towards our data-mining goals.

Project plan

1.Data preparation

Since the kaggle competition states that the data is already pretty clean, there is not much to do in terms of preparation. The small number of rows with missing values can either be dropped or substituted with the average of nearby values. The main preparation will consist of taking a look at the data and getting a general understanding of the most important parameters. Then dropping things like id and such, so they wouldn't affect the outcome of our data-mining.

The estimated time will be upwards of 12 hours, as there isn't too much in need to be done, but a good understanding of the data would vastly improve the efficiency of the following steps.

2.Model development

Firstly implement a quick and easy to understand model such as k-nearest-neighbours to have a good baseline that can learn fast and get some results right away.

Then move on to more complex models such as ARIMA or more compute intensive models such as random forest classifier and clustering.

The estimated time of this task would be about 50 hours of collective development. Giving each member their own models to develop and then discuss.

3.Model evaluation

It is needed to assess the model accuracy using metrics like Mean absolute percentage error which can be gotten from trying to predict the sales of things in test.csv.

Validating is needed to see whether our model performs poorly with already accessible data.

We will need to document the findings during this and find improvements to our models.

The estimated time needed for this task is about 20 hours of collective assessing and improving.

4.Documentation and Reporting

During the development of our project we will be documenting our ideas and the process of which they were implemented .

We will use the results and the documentation to produce the final presentation of what we have learned with this project.

It is estimated to take about 8 hours of our collective time.

Methods and tools include:

- Python with Pandas, Numpy as the primary programming language as it has all the necessary packages for our goals
- Methods such as ARIMA or LSTM for model development, with scikit-learn or tensorflow from python. Other models will probably be used as well to experiment with data and predictions to see what gives the best results.
- Model evaluation will be done with Mean Absolute Percentage Error, as it ensures rigorous evaluation so that the model meets the success criteria.
- Jupyter Notebooks is needed for Documentation and Reporting. As it is familiar to the group members and provides great options for commenting and documenting the written code.

This project plan ensures a systematic approach to the project with some sort of understanding for what needs to be done during this process.