GLOBAL ACADEMY OF MATHEMATICAL AND ECONOMIC SCIENCES

March 2022

DATA SCIENCE PORTFOLIO PROJECT

Probabilistic Sales Forecasting with Discrete Distributions

Project Creator/Supervisor: Zion Pibowei

Course: Probabilistic Modelling

THE PROBLEM

You have been approached by the sales manager of the largest supermarket in the country. The business objective is to understand consumer behaviour and purchasing patterns of the store's customers in the 1st quarter and forecast sales outcomes throughout the 2nd quarter. You've been provided with sales records from Q1. Your objective is to work with your team to create reliable forecasts of sales performance in Q2 and report your findings and recommendations to the stakeholders in terms of quantifiable metrics and actionable insights.

KEY OBJECTIVES

After evaluating the business requirement and understanding the complexity of the business objective, your team has decided to investigate (i) the expected number of sales in Q2 daily, (ii) the distribution of quantities that will be sold across the various branches of the org, (iii) the precise location where the sales will be made, (iv) the precise product line for each sale that will be recorded daily (v) the expected waiting times between each sale, branch by branch.

To approach these objectives comprehensively, probabilistic models are required, not just for predictive accuracy but also quantitative treatment of the likelihood of the predicted outcomes. Your solution will cut across the methods below. **Note:** there are 2 files accompanying this document: a brief run-through of what is expected in the implementations below, and a Jupyter notebook quick-start guide.

1. Binomial Experiments:

(a) Use a binomial experiment to simulate the number of times sales will be recorded in Branch A in Q2 daily. Repeat the simulation for branches B and C, and aggregate the results.

2(b) Simulate the quantity of products that will be sold in each of the branches in Q2 daily. ***Note the difference between this and the above objective!***

2. Multinomial Simulation & Classification:

(a) Run a simple multinomial simulation to predict quantity of sales aggregated by branch and classify the branch where the maximum sales would occur daily in Q2. Repeat the steps for predictions based on product line.

(b) Set up & implement a **multinomial Naive Bayes** algorithm to predict the branch where each aggregated sale predicted by a binomial simulation will be made daily. Implement the algorithm to classify the product line for each sale batch daily.

3. Poisson Processes:

Use the Poisson process to predict the quantity of products that will be sold each day across each of the branches in Q2. Run a different simulation to estimate the waiting times between each sale, from day to day.

4. Robust Bayesian Inference:

Combine the results from the multinomial and Poisson processes in the Q2 simulation for all 3 branches, and use Bayesian inference to predict the product line for each sale that will be made in Q2 given the branches.

THE DATA

The supermarket's IT department has provided the following link for you to access the database containing the sales records for Q1.

https://github.com/gamesconsort/cga-internship-projects/blob/main/probabilistic-sales-forecasting/data/supermarket_sales_data

Don't worry, it's not a database ☺

You are to **ingest this data directly** to your workspace and work with it. *Please, do not try to download to local storage and then import from there!*

The first thing you'd most likely notice is that the data was captured for Q1 2019. To make the project relatable to current times, replace 2019 in the date field with 2022 so that the rest of the project will be a study of the outlook for Q2, 2022.

SCOPE, ASSUMPTIONS & CONSTRAINTS

The focus of this project is on modelling future outcomes using discrete probability distributions. As such, only the discrete and categorical variables of the data should be considered for fitting probabilistic models on the data. The target business variable of primary focus for this project is **Quantity**.

The fundamental assumption to be made for the modelling procedure and probabilistic simulations in this project is that the time series of the relevant variables is (approximately) **stationary**. You are required to make a quantitative or visual justification of this assumption during your EDA.

By assuming stationarity, one can use outcomes at different times in the past to make predictions for different times in the future shifted uniformly from the historical times. The reason is that, for a stationary time series, there is an expected identicalness between outcomes of a variable at one point in time and outcomes of that variable when shifted in time.

Based on this assumption, all your simulations and predictions should be run for a uniform time delta from each day in the historical records. That is, predictions for each time in the future $t + \Delta t$ should be based on each specific time t in the data, where the time delta, Δt, is uniform across all the predictions of future outcomes. Various statistical properties of the data over the entire time history can still however be learned by the probabilistic models and included in making each specific prediction at $t + \Delta t$.

PROJECT WORKFLOW

There's no single specific requirement for how your workflow should be organised, so it's left for your team to strategise on an effective workflow for this project. The following, however, could serve as a guideline and can be adopted/extended if or as you wish.

- **Business Motivation & Requirements:**

  In this section, describe the business problem and explain how and what your team has reasoned about it. Describe any business questions you have narrowed down from the problem. Highlight what you have identified as the business motivation and business goals for solving the problem. For example, it is strongly suggested that you present your understanding of the business domain, how well you understand the outcomes expected by the stakeholders, and why solving this problem is important for them. You can outline specific business requirements that will guide the execution of the project, e.g., client needs and expectations, success metrics, etc.

- **Data Motivation and Objectives:**

  In this section, you are to translate the business problem into objectives that data science techniques can target. The brief outline in the opening paragraph of the **Key Objectives** section can be used as a guideline. After framing the data objectives, specify the motivation for (and advantage of) approaching the business problem as a data problem. Finally, specify the operational requirements for the rest of the workflow, including data sources, data project workflow, proposed modelling approach, model targets, scope and constraints, etc.

- **Data Ingestion & Initial Analysis:**

  This is the beginning of your technical workflow. Load the data into your workspace, perform initial inspection of the data, and carry out any necessary treatment. When writing your report, this section is where you describe your workspace and the tools you used for analysis of the data.

- **EDA & Business Insights:**

  Carry out a comprehensive exploratory data analysis to understand the distributions that underlie the observations in the data, the patterns and dependencies that exist in your data, and extract any useful business insight from the data. Use your EDA to quantitatively or visually justify any assumption that you will be making in your solution.

- **Model Motivation and Methodology:**

  So far you have specified the business and data motivations. Now you need to specify the motivation for your choice of modelling techniques (probabilistic). Describe the specific objectives you aim to achieve with this approach and outline the modelling workflow you intend to follow to tackle the data problem. Give a brief background/description of each of the probabilistic methods you will be using, such as binomial experiments, Poisson processes, etc.

- **Probabilistic Modelling**

  This is the largest component of your workflow that will include 4 key methods in which you will build several models: (i) binomial simulations, (ii) multinomial classification, (iii) Poisson processes, and, (iv) robust Bayesian inference.

- **Summary of Predictions**

  Provide a summary of results from all your simulations.

- **Discussion and Business Recommendations**

  This section can be left for your written report.

PRESENTATION

After completing the project, your team is to come up with a presentation pitching your solution and outlining the steps and significance of your work. Your presentation could be in the form of a written report and, optionally, a recorded slide presentation, although you are encouraged to do both.

Your report should be structured according to the sections of your workflow, and should describe in detail every aspect of your solution from business objectives to modelling to recommendations. However, it should begin with an **Introduction** section in which you give a background about the domain and the nature of the problem, outline the scope and significance of the project, and summarise the steps you took and the results you achieved. This should be followed immediately by the sections **Business Motivation & Requirements** and **Data Motivation & Objectives**.

For the technical aspects of your workflow such as EDA or probabilistic modelling, your report should include relevant visualisations from your solution and interpretation of these visualizations. Code snippets can also be included if they are necessary for an explanation you want to make. Your report should end with a section on discussions and recommendations in which you explain in lay terms what you achieved in your work, including the interpretation and significance of your solution in terms of quantifiable metrics and actionable insights.

Note: Your entire report should read in past tense, indicating that you are discussing a project you have completed.

For the recorded slide presentation, you could follow the structure outlined above, however, your presentation should be simple and compelling. Discussion on any technical aspects of your workflow should include visualisations and preferably be presented from your workspace.

SUBMISSION

This project is to be carried out in teams. There are 2 deliverables to be handed in: your team solution in Jupyter Notebook or Google Colab, and your written report (and/or video presentation). A team repo should be created on GitHub where these can be accessed.

You are to submit this project latest by **Friday, May 13, 2022.**

HACKATHON

A hackathon will be held towards this project from **April 15 – 19, 2022**. The hackathon will focus on the technical aspects of this project, where teams will compete on creating the most impactful solution while employing the best data science and coding best practices. Details on this hackathon will be provided separately.