Data Science 241: AF project

In this project you will apply your statistical and Python programming knowledge to perform data analyses. The course notes of the first term (Topics 1 and 2) contain the theory and examples of all methods required to complete the project. You will be working in groups of three and may choose your own group members. The project will contribute 15% to your final mark.

- **Background**

  Data science and statistical learning have disrupted the traditional approach to retail and e-Commerce (link: Data Science in e-Commerce). Some further examples of data science applications in e-Commerce are listed in this blog (link: blog). The rapid advancements in the internet, user applications and data capturing devices enable modern businesses to gather data at a rate previously deemed impossible. Companies rely on the skills of data scientists and statisticians to wrangle the data and extract meaningful insight to increase profits, predict behaviour or to improve the customer experience.

  In this project, you are given the sales and purchase records of a general online store called **ShopAlot**. **ShopAlot** operates in all nine provinces of South Africa. To make a purchase on **ShopAlot**, customers must create a profile where the company collects the age and province of the customer. The company also gathers historic information on clients such as the age of the user profile, the number of previous orders the client has made, the average amount the client spends per order and internal sales and advertising indexes. Various "in-store" variables are recorded when a client places an order and makes a purchase.

  The lead data scientist at **ShopAlot** believes that some of these variables are related to the amount that the client spends on an order. In this project, you must explore the data and build a model to predict the `Sales` (described below) of a client. The model to predict `Sales` must then be used to interpret which factors influence these variables. To develop a model, you can consider (but are not restricted) to the regression techniques studied in the Data Science 241 module for possible use/implementation.

- **Data Description**

  The dataset contains observations of clients that visited the website. Only clients who made a purchase are considered. Each row/observation refers to a single visit to the online store. The dataset must be used for regression analyses. It may be assumed that the data were collected randomly and independently. The following variables are given in the dataset:

  - `Sales`: This is the dependent variables that indicates the total amount spent (in Rand) on each order.
  - `Month`: Month in which the purchase was made (12 calendar months are given).
  - `AgeInMonths`: The age in months of the client.
  - `ProfileInMonths`: The time in months since the user created the profile.
  - `Subscribed`: A variable indicating whether the client is subscribed to promotional emails. If a client is subscribed, the variable indicates to which promotional email list the client is subscribed to or whether the client is not subscribed.
  - `NumPrevOrders`: The number of previous orders of the customer.
  - `AvgSpend`: The average amount (in Rand) of all previous orders placed by the customer.
  - `POSR`: An internal score (Point-of-Sales Ratio) collected by **ShopAlot**, which indicates the client's likeliness to respond to advertising.
  - `OPR`: An internal score (Order-Purchase Ratio) collected by **ShopAlot**. It represents the ratio of the number of times a client makes a purchase to the number of times the client adds items to a basket.
  - `Online`: The time (in minutes) spent on the website by the customer before logging off.
  - `Discount`: The amount of discount offered on the basket of goods purchased by the client in the order. This value is given as a percentage, i.e., between 0 and 100.
  - `ShippingOption`: A variable indicating which shipping option the client selected.
  - `Platform`: A variable indicating the device (iOS, Android, Computer) used by the client.

- `Payment`: A variable indicating the payment method used by the customer.
- `Ad_1`: This variable indicates if advertisement 1 was used ("Yes") or not ("No"). This is a social media targeted advertisement that displays an image and a short description of a product that the customer might like.
- `Ad_2`: This variable indicates if advertisement 2 was used ("Yes") or not ("No"). Advertisement 2 is only offered on the website for clients viewing products. It displays the popular "frequently bought together" items when a client clicks on certain other products.
- `Ad_3`: This variable indicates if advertisement 3 was used ("Yes") or not ("No"). Advertisement 3 is only offered on the website homepage. This is the "Hot Products" advertisement that shows a list of popular or new products that a client might want to buy.
- `SocialMedia`: This variable indicates which social media channel was used to contact the client in relation to `Ad_1`.
- `ShippingPayment`: A variable indicating if the shipping is Free or Paid for by the client.
- `Category`: Indicates the category of the majority of items in the cart.
- `DayOfWeek`: The day of the week on which the client logon to **ShopAlot**.
- `Province`: The province from where the client logon to **ShopAlot**.

- **Project Criteria and Outcomes**

  It is expected that you use the course notes extensively to perform your analyses. However, you may use techniques not covered in Data Science 241 to analyse the data, especially for the predictive modelling section of this project.

  At the end of the project, you should demonstrate the ability to:

  - process and visualise data using Python.
  - perform the required steps in Python to build a model for regression.
  - perform analyses in Python to investigate the performance of the model.
  - perform analyses in Python to validate the assumptions of the model graphically and statistically.
  - clearly interpret the statistical relationships that are significant in the model.
  - clearly explain how the **ShopAlot** can use the regression model to understand the factors that drive sales.

- **Format and Due date**

  Take note: One group member submits on behalf of the group.
  Date: Sunday, 6 October 2024 (23:59).
  You will be required to submit three documents on SUNLearn:

  - Completed marking rubric indicating for your group project.
  - Typed research report (Word or pdf), adhering to the specific sections as stipulated below.
    * Title page: Suitable project name, group name and group member information (name, surname and SU number). Only include student information on the title page and not on any of the other project pages.
    * Page limit: 10 pages (body of the text: excludes title page, table of contents, reference list).
    * Font size: 12 pt.
    * Line spacing: 1.15.
  - Jupyter notebook with your appropriate Python code to reproduce your results. Include section headings and comments to improve the flow of the notebook.

- **Research report sections**

  - **Problem statement**

    In this section you can give some background on the context of the problem and discuss the aim of the analysis and why it is important (i.e. what will the impact be for **ShopAlot**).

– **Exploratory data analysis**

The first component of any analysis is to investigate all variables (univariately) to understand the data well. This usually entails visualisation and graphic presentation of all variables as well as descriptive statistics which may be presented in tables. This will drive and motivate decisions you will make in the analysis. DO NOT INCLUDE ALL your visualisations in the report, this will be visible in the Jupyter Notebook. The idea is to highlight the most important aspects of the data behaviour. Do not exceed **FIVE** visualisations in this section. Less is more and try to include visualisations with impact and not just simple bar graphs summarising one variable at a time. You will have to interpret each visualisation that you include.

– **Data processing**

In this section give a summary of the changes you have made to the data. Therefore, explain any data coding (categorising or dummy variables) and give motivation for your choices.

– Training and validation sets: To reduce overfitting, you should partition the data into training and validation sets. Decide on an appropriate partition (you may need to research/reference this to ensure it is based on best practice) and split your data into training and validation sets (randomly assigned). The objective is to use the training set to fit/estimate various models and then use the validation set to compare the prediction accuracy of each model, using your preferred measure of goodness of fit or prediction accuracy.

– **Evaluation metrics**

In this section explain which metrics / approaches you will be using to evaluate your models and provide motivation for your choices. Also include a brief explanation of how the measures can be used for interpretation.

– **Linear regression analysis**

A linear regression model must be constructed to interpret which variables are related to sales, and the model must be used to give insights to how **ShopAlot** can increase sales. Please note that some sort of model building is required (*i.e.* you cannot simply present the full model with all possible predictors included). You can use any metric and any model selection algorithm as long as your choice is justified/described. You may even compare the different methods of model selection if you so choose. However, be selective and deliberate about the output you choose to show.

You are free to consider any feature engineering techniques such as, but not limited to, transformations on variables and interactions.

You need to describe each step clearly and concisely in this model-building process. The choice of your final model should be based on both the training and validation sets.

After selecting the best model based on your selection criteria, the model must be used for interpretation. Some possible interpretations include

* Are any of the advertisements effective? If they are, which of the advertisements are effective in increasing sales?
* Are there any weekly/seasonal sales trends?
* Does discount lead to more sales?
* Do certain categories result in more sales?
* Which attributes drive sales?

Note that you should consider transformations on the response and/or predictors, making dummy variables, and/or creating interactions. You may also research more advanced data pre-processing and feature engineering approaches. However, avoid making such a complicated model that you cannot gain any insights into what drives sales at **ShopAlot**. Take note: You may include visualisations in this section if it contributes to your interpretation of the model. Again, it is important to be concise.

- **Predictive Modelling**

  After understanding which factors drive sales, build a predictive model to predict sales as accurately as possible. Here the main objective is prediction—not interpretation. You are expected to spend some time researching other methods/models that can improve the predictive performance of the model built for interpretation. Use the ISLP textbook and its labs as well as resources like Kaggle to research some more advanced statistical learning models. Some examples include generalised additive models, decision trees and random forests, boosting, and neural networks. Do not provide more than two predictive models in the report. You should briefly discuss the method(s) you select and explain how it was implemented, as well as discussing the results obtained.

- **Concluding remarks**

  In this section you will provide a brief overview of the main findings and give your final advice to **ShopAlot**.

- **Marking**:
  - After submitting your final project, you will have to rate the contribution of each group member using a SUNLearn questionnaire by Tuesday (8 October) at 17:00. This rate will be used to adjust the mark of each individual group member.
  - Each student (not group) will be marking two projects by using a provided marking rubric. The marker and group information will only be known to the lecturer, therefore, the review process is anonymous. You have to submit your completed rubrics by Sunday (13 October), 23:59. Everyone starts with a 100% participation mark for the marking component. However, if the marks you allocate are inconsistent with others (standard deviation of approximately 10%) and it is clear from your marking rubric that you did not take the marking exercise seriously, marks will be deducted from your own final mark (up to 10%). The lecturer will be the main marker and will grade every project. However, each project will be marked by three students as well. The average mark obtained from the three student markers will be taken into consideration, but cannot negatively impact the mark you have obtained from the lecturer.
  - A marking rubric will be made available on SUNLearn in due time to assist with the preparation of your project report.
  - Some general remarks:
    - You will be penalised for unnecessary output that makes your submission needlessly long. You should:
      - Not give unnecessary output.
      - Not show data or data frames in the typed report.
      - Not use any methods you do not understand fully and cannot explain/justify.
      - Restrict your discussions in the report to only the essential facts and refrain from giving more description than what is relevant to make your argument.