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 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 25% 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 records of Maties Marketplace—an online retail store. Maties Marketplace operates mostly in South Africa, but recently started engaging with international customers. To make a purchase on Maties Marketplace, customers must create a profile where the company collects the age and country of the customer. The company also gathers historic information on clients such as the age of the profile, the number of previous orders the client has made, the average amount the client spends per order and internal sales and advertising indices. Various "in-store" variables are recorded when a client places an order and makes a purchase. The lead data scientist at Maties Marketplace 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 must then be used to interpret which factors are related to `Sales`. 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 and actually made a purchase. Each row/observation refers to a single sale of the online store. 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 variable that indicates the total amount spent on each order.
  - `Month`: Month in which the sale was made.
  - `AgeInMonths`: The age in months of the customer.
  - `ProfileInMonths`: The time in months since the user created the profile.
  - `Subscribed`: A variable indicating whether the customer is subscribed to promotional emails. If the customer is subscribed, the variable indicates to which promotional email list they are subscribed.
  - `NumPrevOrders`: The number of previous orders of the customer.
  - `AvgSpend`: The average amount (in Rands) of all previous orders placed by the customer.
  - `POSR`: An internal score (Point-of-Sales Ratio) collected by Maties Marketplace, which indicates the customer's likeliness to respond to advertising.
  - `OPR`: An internal score (Order-Purchase Ratio) collected by Maties Marketplace. It represents the ratio of the number of times a customer makes a purchase to the number of times the customer adds items to a cart.
  - `Online`: The time (in minutes) spent on the website by the customer before making a purchase.
  - `Discount`: The amount of discount offered on the cart of goods purchased by the customer in the order. This value is given as a percentage, *i.e.*, between 0 and 100.
  - `ShippingOption`: A variable indicating which shipping option the customer selected.
  - `Shipping`: A variable indicating if the shipping is `Free` or `Paid` for by the customer.
  - `Country`: The country to which the customer's order is shipped.

- `Platform`: A variable indicating the device (iOS, Android, Computer) used by the customer.
- `Payment`: A variable indicating the payment method used by the customer.
- `DayOfWeek`: The day of the week on which the customer makes a purchase on Maties Marketplace.
- `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 customers viewing products. It displays the popular "frequently bought together" items when a customer 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 customer might want to buy.
- `SocialMedia`: This variable indicates which social media channel was used to contact the customer in relation to `Ad_1`.

- **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. The format of the project gives you the freedom to explore different data processing and model building techniques.

  The following criteria will be considered when marking the 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 business can use the regression model to understand the factors that drive sales, or the probability of making a sale.

- **Submission(s) and Project Grading**

  For your final grade for the project, you will need to submit/perform the following:

  - Submit your Jupyter notebook, with your appropriate Python code.
  - Submit a prerecorded 10-minute presentation with accompanying slides, where you clearly motivate your approach and explain your findings.

Your Jupyter notebook and oral presentation will be graded according to the following criteria:

Criteria	Descriptpion	Contribution
Visualisation	All graphics/visualisations provided summarise, describe or extract a **relevant** characteristic or feature about the dataset or input data with thorough discussion. Reasons for presenting these visualisations are clearly defined. Visualisations that significantly enhance your argument or are a summary of similar results are preferred.	10
Metrics	Metrics used to measure performance of a model or result are clearly defined and reported. Metrics are justified based on the characteristics of the problem.	10
Validation / Training	A validation-training set approach has been used.	10
Algorithms & Techniques	Algorithms and techniques (as well as any other methods) used in the project are thoroughly discussed and properly justified based on the characteristics of the data/problem.	10
Implementation	The process for which metrics, algorithms, and techniques were implemented with the data has been thoroughly documented and well-described in your Jupyter notebook and following a logical sequence, justified by the analyses. Jupyter notebook contains clear comments to describe each step of the application.	10
Refinement	The process of improving upon the models generated is clearly documented. All steps in building the final prediction model have been clearly justified and described.	10
Model evaluation & Interpretation	The final model's qualities (such as parameters) are evaluated and interpreted in detail.	10
Creativity	How much creativity, initiative, and ambition did you demonstrate? Did you challenge yourself or did you do only the bare minimum?	15
Overall impresssion & Write-up	How effectively does the write-up communicate the goals, procedures, and results of the study? Are the claims adequately supported? Does the writing style enhance what you are trying to communicate? How well is it edited? Are the statistical claims justified? Are text and analyses effectively interwoven? Are the discussions relevant to the final result/model?	15

Please note: You will be penalised for unnecessary output that makes your submission needlessly long. Although all code needs to be shown you should:

- Not give unnecessary output.
- Not show data or data frames during your oral presentation.
- Not use any methods you do not understand fully and cannot explain/justify in the oral. During the oral presentation, you must discuss the details on all methods used and will be penalised if you cannot explain what you have done and why you have done it.
- Restrict your discussions in your oral presentation to only the essential facts and refrain from giving more description than what is relevant to make your argument.

- **Guidelines for the AF project**

The following guidelines are loosely structured and are simply intended to guide you through the model selection and development process and ultimately arrive at a useful prediction model. There is no single correct way of presenting your answers, but you should motivate your approach adequately. You can analyse the data in any way you feel will give you useful fit/predictions, as long as you describe and correctly motivate the steps in developing your model. Answers between groups can/should differ due to this open-ended approach.

- **Data Exploration and Sample Descriptive Statistics**

  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. For nominal variables this will usually entail something like bar plots, boxplots (as the variables relate to the response) and frequencies (absolute and relative). For continuous/numeric variables you may wish to plot histograms and fit distributions and inspect the means, standard deviations, percentiles, 95% confidence intervals for means etc.

- **Create Training and Validation Sets**

  To reduce over-fitting, 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.

- **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 Maties Marketplace 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, interactions and resampling.

  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.

  You are allowed to use any performance measure to evaluate your model.

  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?
  - Should the company continue with international orders?
  - Are there any weekly/seasonal sales trends?
  - Does discount lead to 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 preprocessing and feature engineering approaches. However, avoid making such a complicated model that you cannot gain any insights into what drives sales at Maties Marketplace.

- **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. In your oral presentation, you should briefly discuss the method you selected (how it works and how you implemented it) and discuss the results obtained by the selected method.