Statistics 244

AF Project 2024

Deadline: 10 October, 15:00

Introduction

In this project you will apply your statistical and `R` programming knowledge to perform data analyses using linear models. The course material contain the theory and examples of all methods required to complete this project. This project will be done in groups for which students are randomly allocated to. Note that students are not allowed to change groups. There are two data sets, and each group must choose one data set on which the project will be based. There is a limit to the number of groups per data set, and group leaders must send their choice of data set to the lecturer. This will work on a first-come-first-serve basis. For the chosen data set a group must determine the best linear regression model, using all methods described in Statistics 244.

Data set 1 - Airbnb

Cape Town, with its stunning landscapes, rich history, and vibrant culture, has become a popular destination for tourists from around the world. Airbnb, a platform that allows travellers to find unique accommodations ranging from cosy apartments to luxurious villas, provides a fascinating lens through which to explore the dynamics of this tourism hotspot. The data set for this project encompasses a variety of Airbnb listings in Cape Town, capturing essential details such as property types, prices, locations, and guest reviews. This data offers a unique opportunity to delve into the factors that determine rental prices. The Cape Town Tourism board has asked of you (a data analyst recruited by Airbnb) to determine which factors influence prices of Airbnb rentals the most in Cape Town. In addition, you must build a predictive model to predict the prices of future rentals based on the provided data.

Data on 7 324 rentals were provided by Airbnb in the `project_airbnb_data.txt` file. Each data entry contains the variables collected on a single rental. The following variables are given in the data set:

- `price`: This is the dependent variable that shows the price of the rental (in Rands).
- `dist`: The distance to the nearest central business district.
- `season`: The season in which the rental occurred.
- `rating`: The review of the tenants normalised to a 0 to 10 scale. 10 indicates an excellent review.
- `size`: The size of the listing in square metres.
- `rooms`: The number of rooms of the listing.
- `bathrooms`: The number of bathrooms of the listing.
- `area`: Location in Cape Town.
- `max_guests`: The maximum number of guests allowed.
- `close_to_beach`: An indicator showing if the listing was near a beach.

Data set 2 - Credit

In the world of finance, credit scoring plays a critical role in determining the creditworthiness of individuals and guiding lending decisions. The data set for this project comes from a leading financial institution and contains detailed information on loan applicants, including demographic data, credit history, income levels, employment status, and education details. This rich data set provides a comprehensive view of the factors that influence credit scores. You were tasked to build a predictive model to predict future scores of loan applicants. In addition, your aim is to uncover key patterns and trends that can help improve the accuracy of credit scoring models, ultimately leading to fairer and more informed lending practices.

The data set, provided in `project_credit_data.txt` consists of 8 160 individuals and each observation consists of a credit score for a client as well as their personal information. The following variables are given in the data set:

- `customerID`: id of the client
- `score`: credit score of the client, a larger score indicates better creditworthiness
- `age`: age of the client in years
- `income`: average income of the client in Rands for the past year
- `education`: education status of the client
- `nrCC`: number of credit cards owned by the client
- `yearsEmpl`: years employed by the most recent employer of the client
- `maritalStatus`: marital status of the client
- `employmentStatus`: employment status of the client
- `history`: number of missed payments by the client

Project Criteria and Outcomes

It is expected that you use the course notes extensively to perform your analyses. You may do additional research to improve the efficiency of your programming, but the models used **should be based only on the work / theoretical models you have covered in Statistics 244**. The format of the project gives you the freedom to explore these different data processing and model building techniques. There are guidelines given below that will guide you through what is expected of you. 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 `R`;
- perform the required steps in `R` to build a linear model for regression;
- perform analyses in `R` to investigate the performance of the model;
- perform analyses in `R` to graphically and statistically validate the assumptions of the linear model and to perform relevant diagnostics;
- clearly interpret the linear statistical relationships that are significant in the model.
- clearly explain how the business can use the regression model to understand the factors related to the response.

Rules

The above mentioned primary outcomes will be considered with the following secondary outcomes, rules, and tips:

- The use of the `ggplot2` package for data visualisation is strongly recommended. Although you are allowed to use `R (base)` plotting tools, credit will be awarded for the use of `ggplot2`.
- Besides for `R (base)`, the only packages allowed in the project are `leaps`, `nortest`, `dplyr`, `ggplot2` (with `scales` and/or `ggpubr`).
- You will be **penalised** for excessively long explanations or for printing out unnecessary information. **Please do not print out entire datasets; rather use `head(dataset_name)`**. Interpretations must be concise and motivated by the model findings.
- In addition to other measures described in the course, this project will make use of root mean square error (RMSE) to compare whether one model fits better than another on the final test set. This will be calculated as $RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} e_i^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$.
- Whether a model is over-fitting the data must always be considered when testing a model (see additional notes on over-fitting). It is advised to use a validation set approach to validate the model performance. This implies that you will randomly need to sample a subset of the observations from a dataset and then use this to define, estimate and validate your model and then determine the RMSE on the remaining observations (*i.e.*, do not calculate the RMSE on the same data as was used to fit the model).
- Ensure regularly that all your code is correctly written in an `R` code chunk within the `Rmd` file. Please note that `R` does not automatically store objects in the `Rmd` file if you type code in the console.

- You should master the ability to work independently within your chosen groups, *i.e.*, without significant help from a lecturer. If required, you must set up a consultation with a lecturer to ask questions related to the project. The lecturer will not respond to regular, short questions asked via email. Furthermore, the lecturer has the right to restrict further consultations with specific groups to ensure that all groups are treated fairly.
- There is no single correct answer for all the questions. Emphasis must be placed on the motivation of a particular approach. You can analyse the data in any way you feel will give you the best fit/predictions, as long as you describe and correctly motivate the steps in developing your model.
- Students must be able to clearly motivate their approach and explain their findings in the submitted document.

Submission(s) and Project Grading

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

Submission	Description	Contribution to Grade
PDF File	Knitted version of your RMD file (including code chunks, selective output and motivations/explanations in markdown).	60%
Presentation	Each group must create a video of a Power point presentation that summarises their results.	20%
Competition	Submission of your predictions from your final model for a "test" set.	10%
Group Involvement Rating	Each member of the group will be required to rate the relative effort of all other members in the group.	10%

These are further explained in the sections that follow.

PDF file Submission

Your project needs to be completed as an R Markdown file (.RMD). All questions (including code, *selective output* and explanations/motivations in plain text and/or LaTeX equations) should be answered in the file. Once you have completed your project you will need to `knit` your file to PDF format and then submit only this last knitted version. Knit your document regularly throughout the project and not only once at the end, to allow time to trouble-shoot potential problems.

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

- **NOT** print unnecessary output to the console. Only output (*preferably summarised in some way*) that is relevant to your model development steps should be displayed.
- **NOT** print dataframes to the console. Use either the `head()` or `tail()` functions to print only the first/last few observations in a dataframe to the console, and only if it bolsters your motivation in some way.
- **NOT** use any methods (beyond those presented in the course notes) you do not understand fully and cannot explain/justify clearly. You will be penalised if you do not motivate and explain each step of your application.
- **Restrict** your discussions to only the essential facts and refrain from giving more description than what is relevant to make your argument.
- **Restrict** your overall (knitted) project to **10 pages** or less. A severe penalty will be given to groups who submit PDF files of more than 10 pages.

This component will be graded by the lecturer according to the table given on the final page of this document. Note that the scores in the table sum to a total of 80 Therefore, the mark for the `RMD` submission of the project will contribute to your final project mark, as follows:

$$Contribution\ to\ Final\ Project = \frac{Your\ Group\ Mark}{80} \times Weight\ (60\%)$$

.

Video presentation

Each group must submit a video of no more than 5 minutes along with their final knitted PDF submission. During the theory lectures on the 14th and 21st of October these videos will be played for the entire class. The video will contribute in total 20% to the final mark of the project of which 50/ will be graded by the lecturer, and 50% will be graded by the class (an average grade made by the groups).

Competition

Each group will have the opportunity to make three submissions of their final model against the benchmark test set. Each submission must be made by 15:00 by the following dates: 25 September, 30 September, 4 October. These submissions must use the template that will be provided on SUNLearn.

At the end of the project, your group ranking will contribute to your final project mark, as follows:

$$Competition\ contribution\ to\ Final\ Project = Your\ Group\ Score \times competition\ Weight\ (10\%)$$

Your group score is calculated using your RMSE obtained against the test set. Then, let $RMSE_{min}$ be the best possible $RMSE$ for the test data, and let $RMSE_{max}$ be the worst $RMSE$ for the test data. These two boundaries are set by the lecturer. Further, let the group score for the $RMSE$ on the ith submission be $RMSE_i$. Your group score on the ith submission is calculated with

$$Group\ Score_i = 1 - \frac{RMSE_i - RMSE_{min}}{RMSE_{max} - RMSE_{min}}.$$

The final group score will be the minimum of the three submissions. The final score represents a percentage and cannot be greater than 100%, *i.e.*, if your $RMSE$ is smaller than $RMSE_{min}$, you will receive 100% for this part of the project. If your $RMSE$ is greater than $RMSE_{max}$, you will receive zero for this component of the project.

Group Involvement Rating

To ensure that each member of the group contributes equally to the project, each member will be required to rate the their own relative contribution as well as that of the other group members (in a subjective estimate of the percentage contribution to the total project). This will be completed individually and other group members will not be aware of the ratings received from their group peers.

Questions and Instructions

The broad questions for the regression part of the Data Science project are given below. These questions are loosely structured and are simply intended to guide you through the model selection and development process and ultimately arrive at the best prediction model. Answers between groups can/should differ due to this open-ended approach. There is no single correct way of presenting the answers, but you should motivate your approach adequately.

Step 0: Data Exploration and Sample Descriptive Statistics

The first component of any analysis is to investigate all variables (univariately). This usually entails visualisation of all variables as well as descriptive statistics. For nominal variables this will usually entail something like bar plots, box-plots (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. The ultimate goal is to understand your data well.

Even though you should do this prior to beginning the model-building steps, this should not form part of your final output. To do these analyses in your RMD file, but not include the results in the final document (if you prefer to do it this way) you can use `{r, include=FALSE}` in the code chunk header.

Step 1: Create Training and Validation Sets

As previously mentioned, to reduce over-fitting, we would like to 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. So in this first step, create your training and validation sets. **Do not** print/output the complete data frames, but do describe (in a paragraph) what you have done as well as the characteristics (e.g. dimensions) of the datasets.

Step 2: Multiple Linear Regression

Using only the techniques available to you in Statistics 244 you need to find a multiple linear regression model that is best at modelling and predicting the response. 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 presented in the course 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, since dumping all results (including intermediate and unnecessary results) to the document will result in severe penalties.

You are free to consider transformations on variables, if supported by the results. In terms of model complexity, you can limit the model to include only second-order terms (or lower) and first-order interactions, if applicable. You need to clearly and concisely describe each step in this model-building process. The choice of your final model should be based on both the training and validation sets. In addition to any other metrics used, you should compare the performance of your best fitting-model using RMSE on both the training an validation sets. Realise that building a statistical model is a process that may change over time. This may be informed by the client, employer or end-user and they may suggest alterations to the models. However, in this project you may use the competition platform to gauge the performance of your model relative to those of your peers.

Hints

- Use a random seed when creating your training and validation sets. If you do not specify a random seed, the results cannot be reproduced.
- Explain clearly which variables are related to the response variable. If all the variables are not related to response variable, how can the company reduce the amount of data they collect. This will have an implication on the cost of data collection and storage.
- If a categorical variable has too many groups, consider grouping some categories together in a group `Other`.
- Regularly knit and save your project.
- Each group member should work on all parts of the project. Do not divide the project into different, individual sections. Each part of the project builds on the previous sections. Rather, each person should do each section, and the best version as chosen by the group should be used for the submission.
- Start early and work together. Working in a team requires continuous communication between the group members.
- Enjoy the project!

Grading Rubric for PDF Submission

Criteria	Description	Contribution to PDF Submission
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 do not contribute meaningfully to the description will be penalised. 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.	5
Validation / Training	A validation-training set approach has been used.	5
Algorithms and 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	5
Implementation	The process for which metrics, algorithms, and techniques were implemented with the data has been thoroughly documented and well-described in markdown and following a logical sequence, justified by the analyses.	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 and the final model's qualities — such as parameters — are evaluated and interpreted in detail. Some type of analysis is used to validate the robustness of the model.	15
Creativity	How much creativity, initiative, and ambition did the group demonstrate? Did the group submit creative ideas / used critical thinking, or did the group do only the bare minimum?	10
Overall Impression and 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?	10