STAT 562 Project Description – Fall 2023

Due date: Thursday, Dec 12th, by 11:59 p.m.

In this project, you are expected to perform a complete analysis of a dataset. You will submit a written exposition that thoroughly describes your approaches. This is an opportunity for you to amalgamate all your knowledge in machine learning and apply it to a real problem. An important part of modeling is clearly communicating it to others in writing and the written project is how you disseminate your work.

Introduction for the Dataset:

A tour & travels company is offering travel insurance package to their customers. The new insurance package also includes covid cover. The company wants to know which customers would be interested to buy it based on their database history. The insurance was offered to some of the customers in 2019 and the given data has been extracted from the performance/sales of the package during that period. The data is provided for almost 2000 of its previous customers and the goal is to build a model that can predict if the customer will be interested to buy the travel insurance package. I've withheld 100 observations to see how your model performs. **You will submit your final prediction on these 100 observations in the provided csv file (TravelInsuranceTest.csv) along with your project paper.**

A: Target Variable/Label

TravelInsurance - Did the customer buy travel insurance package.

B: Predictor Variables/Features

Age - Age of the customer
Employment Type - The sector in which customer is employed
GraduateOrNot - Whether the customer is college graduate or not
AnnualIncome - The yearly income of the customer
FamilyMembers - Number of members in customer's family
ChronicDiseases - Whether the customer suffers from any major disease
FrequentFlyer - Derived data based on customer's history of booking air tickets on at least 4 different instances in the last 2 years
EverTravelledAbroad - Has the customer ever travelled to a foreign country.

Guidelines for Modeling and Forecasting:

The aim is to build a few machine learning models that can predict if the customer will be interested in buying the travel insurance package based on certain parameters.
It is important to keep in mind the steps you should consider in machine learning:

1. Before we create a model, do some data cleaning, feature selection and exploratory data analysis.
2. Come up with a set of candidate methods that is suitable for the data.
3. Fit the models with training data.
4. Reduce the dimension of features by performing feature selection or dimension reduction.
5. Adjust the tuning parameters using cross-validation or model performance criteria such as error rate, AUC, etc.
5. Check the adequacy of the model fits and possibly revise the model.
6. Compare the models and choose your final model base on the prediction accuracy on the test data.

Outline of the Written Project Paper:

• **Title page and abstract:** You must prepare a title page with an appropriate title and abstract. The abstract should go on the title page. An abstract is a very high-level written summary of the entire project. Main points and results only.
• **Introduction:** This part introduces the reader to the dataset and the analysis method that you are going to perform. This should be written at a very basic level. Remember your reader may not know anything about the area in which you are writing.
• **Model specification:** In this section, you want to describe, in clear detail, the statistical analysis used to specify your candidate models. Pretend as if you are taking the reader by the hand and leading him or her through your thought process which leads to your model selections.
• **Fitting and Diagnostics:** This part of the project should describe the model fitting and diagnostics techniques you used, with the goal of identifying a "final". Identify also what possible deficiencies your final model has. Remember, no model is perfect.
• **Discussion:** Here you want to offer a summary of what you did in the project and present your result in terms of training and test errors. Also, it is a good idea to discuss here other issues related to the data analysis. For example, does your analysis have any shortcomings or lack of generalizability? What were the main problems you encountered? It is OK if your final model is not perfect. Real life data analysis is often more difficult than textbook problems.
• **Appendices (optional):** Use appendices to catalogue extra graphics/plots/output. Basically, I use an appendix to house information that I want the reader to have access to, but feel that it would interrupt the flow of the main body of the paper.

General Advice:

• Please make the formatting such that the report is easy to read. Break your report into sections. Each section should have a title. Use subsections if necessary.
• Integrate R graphics and output into the written text as you see fit. For example, if you want to show the time series itself, embed it into the written work. Look at the style of the way graphics are embedded in your textbook as a guide
• Edit and proof read your written project. You may get other people to read your project and offer comments/feedback.
• I don't have a specific target number of pages for the whole report. You should do enough to provide a full analysis of this dataset, with attention paid to each of the sections listed above. You should end up with around 8-10 pages.
• Have fun!

Grading Scale: (100 points total)

Writing (30 points): Organized, clearly written, comprehensible, model interpretation and grammatically correct.
Analysis (50 points): Were the chosen models, graphs, and data analyses appropriate for the problem? Were the analyses carried out correctly? Were your conclusions about the data set sensible and clearly justified by numerical or graphical evidence? Have considered aspects of how the model can be performed?
Prediction (20 points): How your model preforms on prediction of the unseen data (the 100 examples withheld).