MIS 315 The Final Project

This problem involves Fifa 18 Ultimate Team dataset. The dataset is obtained from https://www.kaggle.com/kevinmh/fifa-18-more-complete-player-dataset. The original dataset is very large. I delete some columns to obtain a more compact dataset. Your main aim is to predict the value of players (eur_value). First two columns (ID, name) represent identification of players. Third column is your response/dependent variable, the value of players (eur_value). Other values are predictors/independent variables. You can obtain more information about predictors from https://www.fifauteam.com/fifa-18-attributes-guide/, https://www.fifplay.com/encyclopedia/player-attributes/ and other similar websites.

a) First drop all goalkeepers from the dataset. If a player has a gk attribute, then it means he is a goalkeeper. **5 points**

b) Analyze your predictors/independent variables using techniques shown in the class. **5 points**

c) Take the first 1000 observation as the training set and use the rest as the test set. **5 points**

d) Fit a multiple regression model to predict "eur_value". Use only the training set to fit the regression model. **5 points**

e) Analyze your estimated regression models. Comment on coefficients, adjusted R square and F statistic of the model. **5 points**

f) Predict "eur_value" in the test set using the regression model obtained in (d). Calculate the mean square error of the test set (MSE). **5 points**

g) Fit a Ridge model and a Lasso model to predict "eur_value". Use only the training set to fit these regression models. Determine the lambda parameter using cross-validation. **5 points**

h) Analyze your Lasso Model. Compare your Lasso Model with the multiple regression model estimated in (d). **10 points**

i) Predict "eur_value" in the test set using the Ridge model and the Lasso model obtained in (g). Calculate MSEs of these models only using the test set. **5 points**

j) Fit a regression tree to predict “eur_value”. Use only the training set to fit the regression model. Determine the number of terminal nodes using cross-validation. **10 points**

k) Predict “eur_value” in the test set using the regression tree model obtained in (j). Calculate the MSEs of the regression tree only using the test set. **10 points**

l) Fit random forests to predict “eur_value”. Use only the training set to fit the regression model. Determine the number of variables used in each split using the cross-validation. Grow 500 trees for random forest. **10 points**

m) According to random forests, which variables are import? Comment. **10 points**

n) Predict “eur_value” in the test set using the random forest model obtained in (l). Calculate the MSEs of the random forest only using the test set. **5 points**

o) Compare MSEs obtained in (f), (i), (k) and (n). **5 points**

p) Create a new categorical variable from “eur_value” called “value_level”. If the “eur_value” is greater than 10,050,000 coded as “high_earner”, otherwise coded as “low_earner”. As a result, “value_level” variable will only include two types of categorical data: “high_earner” and “low_earner”. **5 points**

q) Choose a new train and test set by randomizing data set. %50 of data would be train set and the rest would be test set. **5 points**

r) Fit a logistic regression to predict “value_level”. Use only the training set to fit the classification model. Calculate the accuracy of the model using the test set. **5 points**

s) Fit a LDA to predict “value_level”. Use only the training set to fit the classification model. Calculate the accuracy of the model using the test set. **5 points**

t) Fit a QDA to predict “value_level”. Use only the training set to fit the classification model. Calculate the accuracy of the model using the test set. **5 points**

u) Fit a classification tree to predict “value_level”. Use only the training set to fit the classification. Determine the number of terminal nodes using cross-validation. Calculate the accuracy of the model using the test set. **5 points**

v) Fit bagged trees to predict “value_level”. Use only the training set to fit the classification model. Determine the number of trees used in bagged trees using cross-validation. Calculate the accuracy of the model using the test set. **5 points**

w) Fit random forests to predict “value_level”. Use only the training set to fit the classification model. Determine the number of variables used in each split using the cross-validation. Grow 500 trees for random forests. **5 points**

x) According to the random forest, which variables are import? Comment and compare with the result obtained in (m). **5 points**

y) Compare accuracies obtained in (r), (s), (t), (u), and (v). **5 points**

Rules for the Final Project:

- You will first present your results in front of the class. The language of presentation is English. The last date for sending presentations is 2 January 2021 12:00.
- After receiving feedback from your presentation, you will submit your final project paper and your R code. The last date for sending the final project paper and the R code is 20 January 2021 23:59. You will also make a final presentation on 21st January of 2021 at 19:00. Without that final presentation, you will get zero points. In the final presentation, I expect you to explain your code line by line. If you can not do this, you will get zero points.
- You can form groups up to 3 people. It is expected that students will manage their groups so that everyone performs a fair share of the work, and that all perspectives are heard and considered. The students should be responsive to group communication and get along well with their group members. There will be also a peer evaluation of the group members which might in some cases impact the homework grade.
- Students without a group will complete tasks up to (o). Groups will complete all tasks.
- Only group members can exchange code with each other. Otherwise, no student can exchange code with each other. If you don’t follow this rule, both the student that shares the code and the student that uses the other student’s code will be punished.

General Required Format for the Final Project Paper

Your paper should look professional. The format of the paper is as important as your analysis in the paper. I will deduct points if your paper does not adhere to the required format described below.

- The paper may consist of the following parts: 1) Title page 2) Abstract/Summary 3) Introduction 4) Data and Analysis 5) Conclusion 6) References 7) Tables and Graphs
- You should include a title or cover page including the following: i) project title, ii) group number, iii) names of all group members.

- After the title page include a short summary or abstract summarizing your basic findings. This abstract shall not exceed 250 words.
- In the introduction explain and motivate your research project. Discuss expected results. The reader should get the big picture about your project after reading the introduction.
- Data and analysis part gives the detailed procedures and the discussion of the project. This part can consist of several sections and subsections. These sections/subsections should be labeled with brief descriptions of your quantitative and qualitative analysis. The flow in this section should be linear, and the sections should be well connected to each other in a meaningful way.
- Conclusion is a summary of your results with implications for future research. Keep it brief but compact.
- References section should include full citations of the articles and books you referred to in your write-up.
- All tables/spreadsheets and figures/graphs should be labeled with a title and a description of the analysis (legend).
- Do not use acronyms and abbreviations of variable names in Tables and Pictures. While we use those to represent concepts in our course they are not acceptable in a formal research report. Bear in mind that readers with no Modeling background should be able to read and understand your report.
- Tables and figures should fit on one page or be separated in a logical manner. They should be visually appealing, easy to read and incorporated into the text. Avoid excessive and distractive use of color in tables and graphs. Alternatively, you may put all tables and figures into the appendix. Make sure you refer to each table and figure within the text.
- Paginate the paper.
- Check for spelling and grammar mistakes.