Final Individual Project

Due: November 22nd at 11:59 pm on Canvas

Total Points: 115 points

Submission Requirements: a written report with the required elements/sections in a word or pdf document and your R code.

Description:
The purpose of this individual final project is to get your 'hands dirty' on a real data set and to go through the entire process of a simple data science project from exploring the data, training your data using different techniques, testing your data, and writing a report that can be delivered and understood by other people. It is a fully-written report with comprehensive sentences, not bulletpoints.

You can choose one of the following two Kaggle data science competitions. (You don't need to actually participate in the competition)

- https://www.kaggle.com/c/titanic/overview
- https://www.kaggle.com/c/house-prices-advanced-regression-techniques/overview

The report itself is worth 100 points (details below). There is an addition 10 points for documenting your progresses by submitting two deliverables by their respective due dates (detail below). There is an additional 5 points for your R code. Your R codes should be reproducible and have appropriate comments that document what you are doing. The total project is 115 points in total.

Option: group vs. individual

You have the option of completing this project in a group of two. Each group should make one report and one R code for the project, and members of a group will receive one score for a given submission. Be specific. Make sure your final report and especially your R code are cohesive and consistent. At the end of your report, include a separate section that documents the job division – who did what.

Please send me an email to tell me if you would like to form a group and with whom by Nov 8th.

Documenting Progresses (10 points)

- If you submit the following two things by their due date, you will receive 5 points for each. There will be no extension for these.

    - 1) Introduction Section **by Oct 25th at 11:59pm**
    - 2) Explanatory Data Analysis Section **by November 8th at 11:59pm**
- The purpose to make sure you are working on the project in your free time, instead of working on it in the last minute
- I will just check if you have something intelligent written. I won't actually grade them. But I may leave comments. So, this will also be a great time for you to ask me questions and get some feedback

The report should have at least the following 5 sections:

Introduction (5 points).

- Describe, in your own words, the data set and competition you selected. For example, what is the goal of this competition, what data were collected, why are they collected, etc. … Provide a summary of your story.
- Describe how you would define and measure the outcomes from the dataset. How would you measure the effectiveness of a good prediction algorithm or clustering algorithm?

Explanatory Data Analysis (30 points total)

- [15 points] Choose **at least 5 attributes** (variables) to include in your explanatory data analysis section. You can do a mix of the following two, with **at least three** graphs for data visualization.
    - Give simple, appropriate statistics (range, mode, mean, median, variance, counts, etc.) for the most important attributes and describe what they mean or if you found something interesting. **Note**: You can also use data from other sources for comparison.
    - Data visualization: make sure you create professional graphs with the appropriate **labeling** and **footnotes**. **Important**: Provide an interpretation for each chart. Explain for each attribute why you chose the used visualization
    - Make sure you analyze your results from the simple statistics or the graphs.
- [15 points] Explore relationships between attributes: Look at the attributes via scatter plots, correlation, cross-tabulation, group-wise averages, etc. as appropriate.
    - This can include relationships between an independent variable and your outcome variable (the variable you are trying to predict)

Features Management/Data Cleaning (20 points):

- [10 points] Verify data quality: Are there missing values? Duplicate data? Outliers? Are those mistakes? How do you deal with these problems?
    - Are there a lot of missing values for a certain variable? If so, what should you do? Should you do imputation? Should you include that variable in your models?
- [10 points] Data transformation:

- Are there any data that is highly skewed? What kind of transformation should you do?
- Should you convert certain numerical variables into categorical variables so that you can use the group as a predicator? Why do you do it? How would this help with your predictions?
    - For example, group a numerical variable age into: kids, teenager, adult, senior, etc…

Learning Algorithm Training (35 points total)

- [5 points] Describe the machine learning methods you decide to use in your data with a general description, pros and cons. In the appendix, I provided an example of two methods. I used a table. Feel free to not use a table.
- [30 points] Choose at least 2 machine learning methods. From what we learned in class, you should be able to apply at least 2 machine learning methods. Each method is 15 points. For each machine learning method you choose, include a separate section. Within each section:
    - provide a more detailed description of what you did: the steps you took, the parameter values you used, the function you used, any specific R package you used.
    - After you decide on a model with the features/variables you decide to include, test your model on the test data set provided by Kaggle. Provide any appropriate plots for your results
- For any additional method you choose, there will be a bonus of 15 points.
    - To earn these extra points, you may have to do some learning on your own.

Conclusion (5 points total)

- Summarize your problem, your solutions and your results, etc.

Appendix: an example of the method descriptions

Model	Description	Pros	Cons
Xgboost Regressor	It is an implementation of gradient boosted decision trees. Boosting is a machine learning technique where new models are added to correct the errors made by the existing models until no further improvements can be made.	Can do parallel processing. Can work on very large data set	Likely to cause over-fitting
Extra-tree Regressor	It stands for **Ext**remely **ra**ndomized **tree**s. It randomized the process of tree-building. It selects a cut-point at random, unlike random forests, which tries to find an optimal cut-point for each feature at each node.	More productive when the number of numerical features is large.	Likely to cause overfitting. Computationally expensive.