BAN 502 Course Project Description

Project Description

For this project you are provided a dataset related to the prediction of product failures. Your ultimate objective will be to develop predictive models to predict this response variable. The project features two phases. Each phase is described below. The project will be conducted as a Kaggle competition. You will compete with your classmates to develop the best predictive model that you can.

You can use the link below to access the Kaggle competition: https://www.kaggle.com/t/5008789507b3466e9755c60e091dd4c5. You will need to create a free Kaggle account if you do not already have one.

Phase 1 Project Description

For Phase 1 you will conduct a thorough exploratory/descriptive analysis of the dataset. Please **DO NOT** build any predictive models (e.g., logistic regression, trees, etc.) in this phase.

Assume that your "audience" for this work are non-technical decision-makers.

Phase 1 Deliverables:

There are two deliverables for Phase 1.

Deliverable 1**:** A PowerPoint presentation summarizing your findings from Phase 1. The presentation should be no more than seven slides (including a title slide). Your findings should indicate which variables may be strong predictors of the "failure" variable (this is your response variable) as well as any other interesting descriptive findings. You should include a charts/visuals in the presentation. There should NO VISIBLE R CODE in this deliverable. As noted above, you should assume that the target audience for the deliverable is relatively "non-technical." NOTE: If you create any variables (i.e., by combining or modifying existing variables), please note this.

Deliverable 2**:** A knitted Word document of your Phase 1 R work.

Submit the deliverables via Canvas. THERE IS NO KAGGLE SUBMISSION REQUIRED FOR THIS PHASE!

Phase 2 Description

In Phase 2 you will build predictive models to predict the variable "failure". You will develop multiple predictive models to predict this variable. You should fully document (in your R Markdown file, not in your PowerPoint deliverable) all model building efforts. You should use a training/testing split and may choose to apply k-fold cross-validation when building your model on the training set. Please employ multiple techniques (logistic regression, classification trees, random forests, etc.).

As in Phase 1, assume that your "audience" for this work are non-technical.

Phase 2 Deliverables:

There are three deliverables for Phase 2:

Deliverable 1: A PowerPoint presentation summarizing your findings from Phase 2. The presentation should be no more than seven slides (including a title slide). Your findings should focus on the practical implications of your findings. If your findings are "weak", you should indicate so. You should include appropriate charts/visuals in the presentation. There should NO VISIBLE R CODE in this deliverable. As noted above, you should assume that the target audience for the deliverable is relatively "non-technical."

Deliverable 2: A knitted Word document of your Phase 2 R work.

Submit Deliverables 1 and 2 via Canvas.

Deliverable 3: You **must** submit your model predictions on the "test.csv" file to Kaggle. For each row in the "test.csv" file you should predict whether or not the product will fail (No or Yes). See the "sample_submission.csv" file for an example of how to format your predictions on the "test.csv" file. You can submit multiple submissions if you wish. The submission that performs best will be the submission that

Hints/Suggestions/Warnings for Phase 2:

- Provide a simple summary table showing your models' performance on the training and testing sets.
- There may be missingness that needs to be dealt with.