Project-related PDFs

Project Report.pdf

- In this PDF, there is thoroughly written and detailed information regarding our project along with a bibliography.

Project Presentation.pdf

- This PDF of slides is a more concise version of our Project Report with an executive summary and overview of methodology.

Folders

RMDs and Rendered PDFs

In this folder, there are 2 subfolders: one for modeling and one for EDA and feature engineering tasks

- EDA and Feature Engineering (Folder)
    - *eda_engineering.Rmd*
        - PDF rendered version: *eda_engineering.pdf*
        - EDA on all columns from the original kaggle dataset
        - Feature engineering on all the columns. Produces datasets that are found in the Dataset folder.
- Modeling (Folder)
    - In this folder, there are three Rmd files because the modeling portion of this project was split among three team members and the three PDF files are the respective PDF rendered versions of its Rmd file.
    - There is also 1 Rmd file on the feature engineering portion of this project with its PDF rendered version.
    - Here is a list of the models each file was responsible for generating:
        - *csp571-tiffmodeling.Rmd*
            - PDF rendered version: *csp571-tiffmodeling.pdf*
            - Chi Squared Test on Logistic Regression model to find significant attributes
            - Stochastic Gradient Boosting model
            - Random Forest Classifier model
            - Classification Tree model
        - *alisha_modeling.Rmd*
            - PDF rendered version: *alisha_modeling.pdf*
            - Naive Bayes Classifier
            - Support Vector Machine Classifier
            - K-nearest neighbor Classifier
        - *anette_modeling.Rmd*
            - PDF rendered version: *anette_modeling.pdf*
            - Logistic Regression model

Datasets

In this folder, there are 2 subfolders for the training and testing datasets and a separate .csv file for the output from our Chi Squared test.

- Train (Folder)
    - *joint_numeric.csv* - training dataset with only numeric attributes
    - *joint.csv* - training dataset with all attributes from feature engineering and all original attributes
- Test (Folder)

    i. *joint_test_numeric.csv* - testing dataset with only numeric attributes

    ii. *joint_test.csv* - training dataset with all attributes from feature engineering and all original attributes
- *chisq_test.csv* - output of Chi-squared test on all attributes to subset dataset with only significant attributes
- fake_job_postings.csv - our original dataset from Kaggle