**Machine Learning: Project Code**

Craig Shaffer

December 10, 2022

**File Description**

The file descriptions for project code are displayed below:

|  |  |
| --- | --- |
| **File Name** | **Description** |
| games.csv | Raw game data sourced from GitHub repository nfldata by nflverse. |
| DVOA\_2011\_2021.csv | DVOA data for teams from 2011-2021. Sourced from FootballOusiders.com. |
| CleaningData.R | R file for creating final data set NFL\_2011\_21.csv. This combines data from the two csv files (games.csv and DVOA\_2011\_2021.csv). It also creates new variables and cleans the data of missing values. |
| NFL\_2011\_21.csv | A clean data set to be uploaded to S3 bucket and used for analysis. |
| ExploritoryAnalysis.ipynb | Uses the NFL\_2011\_21.csv file to explore variables that may be useful for predicting the winner of a football game. |
| VariableImportance.ipynb | Uses the NFL\_2011\_21.csv file to determine the top 10 variables for predicting the winner of a game. It then creates a file called games\_Final.csv that contains the top 10 predictor variables and the variable I want to predict (home\_win) |
| games\_Final.csv | Final data to be used for model generation. Contains the 10 most important variables and home\_win. |
| EvaluateModels.py | To prevent Kernel issues, the models had to be run in the terminal. This .py file was created to run and evaluate several models with many hyperparameter combinations. It then generates and exports the results of each model into a csv file. |
| lr\_results.csv | Logistic Regression model results |
| svm\_results.csv | Support Vector Machine model results |
| dt\_results.csv | Decision Tree model results |
| rf\_results.csv | Random Forest model results |
| ada\_results.csv | AdaBoost model results |
| gb\_results.csv | Gradient Boosting model results |
| ReportVisuals.ipynb | Using the data from the result csvs, this file identifies the best model of each type and creates visuals used in the final report. |