Statistical and Predictive Modeling for Analytics II (DATA 2204)
Final Project (25% of Final Grade)
Professor: Ritwick Dutta

Background

Mr. John Hughes has been reviewing the **wireless_churn.csv** and would like you to create a ***three (3) forecasting models*** using Logistical Regression, Naïve Bayes and Voting Ensemble.

Dataset contains: 3,333 observations and 11 variables:

Independent Variables

AccountWeeks - number of weeks customer has had active account

ContractRenewal - 1 if customer recently renewed contract, 0 if not

DataPlan - 1 if customer has data plan, 0 if not

DataUsage - gigabytes of monthly data usage

CustServCalls - number of calls into customer service

DayMins - average daytime minutes per month

DayCalls - average number of daytime calls

MonthlyCharge - average monthly bill

OverageFee - largest overage fee in last 12 months

RoamMins – average roaming minutes per month

Dependent Variables

Churn - 1 if customer cancelled service, 0 if not

The Ask:

Using Python and Jupyter Notebook create the following script:

1. Exploratory Analysis
   a) Conduct **Exploratory Data Analysis (EDA)** using pandas-profiling to help identify key insights from the dataset.

2. Remove Anomalies
   a) Remove outliers using **Isolation Forest**

3. Create Learning Curves for both algorithms (Logical Regression and Naïve Bayes).
   a) Please use recall for your scoring (i.e. scoring='recall_weighted')
   b) Logistical Regression (**solver='lbfgs', class_weight='balanced', max_iter=1000, random_state=100**)

4. Create Optimize models (including ROC/AUC Curves) using the following two (2) algorithms to predict the proper label classification:
   a) Logistical Regression (**solver='lbfgs', class_weight='balanced', max_iter=1000, random_state=100**)

   b) Naïve Bayes

Note: You don't need to create 'Original Models', just Optimized Models

5. Create **one (1)** Ensemble Voting Model, to predict the proper classification, which includes:
   a) **one (1) algorithm** (i.e., Logistical Regression or Naïve Bayes)

   b) **one (1) Bagging or Boosting Technique** (Bagging, Adaboost, or Gradient Boosting)

6. Next Steps:
   a) **Identify (1) algorithm** you created (i.e. Logistical Regression, Naïve Bayes, or Voting Ensemble) that should be implemented by Mr. John Hughes.
   b) **Identify and justify two (2)** next steps that could be used to help enhance the usability of the model you chose.

Final Documents

1. A PowerPoint deck (**PPT or PPTX**) to report your analysis, findings, and conclusions. **See Appendix A for details**

Random State = 100 for all sections

Note: Please ensure that all key facts are in your slides and not in the notes section

2. Python code using Jupyter Notebook (in .html)

Appendix A

PowerPoint Requirements:

Cover Slide

- Title: Final Project (DATA 2204)
- Name (First and Last)
- Student Number

Slide 1 ***(1%)***

- Problem statement (i.e. the ask from Mr. John Hughes)

Slide 2-6 ***(3%)***

- Using exploratory data analysis (EDA), identify and explain **three (3) key insights** from the **UCI_Dataset.csv** dataset from the Pandas Profiling report (i.e., please don't use summary page).

Slide 7-8 ***(4%)***

- Present the Learning Curves for both algorithms and explain **two (2) key insights for each associated Learning Curve**. ***Total of four (4) key insights are required.***

Slides 9-16 ***(12%)***

- Present the Classification Report and ROC/AUC of each of the optimized models (i.e. Logistical Regression and Naïve Bayes) and Explain **three (3) key insights for each optimized model** (i.e., Precision, Recall, F1, Support for both summary and detailed metrics). ***Total of six (6) key insights are required.***

Slide 16-18 ***(3%)***

- Present and Explain the results of the Ensemble Voting model and how it compares to the other two optimized models (Logistical Regression and Naïve Bayes).

Slide 19-20 ***(2%)***

- Identify **one (1) model** that you created (i.e. Logistical Regression, Naïve Bayes, or Voting Ensemble) that should be implemented by Mr. John Hughes.
- Identify and justify **two (2) next steps** that could be used to help enhance the usability of the model you chose.

Code Requirements:

Python code using Jupyter Notebook in HTML (.html) format. **Note: 50% Penalty for missing Jupyter Notebook HTML file**

NOTE: The number of slides is a guideline not a requirement

Please post your PowerPoint (.ppt or .pptx) and HTML (.html) Jupyter Notebook under Final Project by Wednesday, August 14th, 2024 @ 11:59 p.m.

HINT: Use Week9e-Tutorial-IsoForest as your starting point

Note: 50% Penalty for missing Jupyter Notebook HTML file