Final Project: Programming Option

Overview

Prepare eminently readable, enhanced and updated Jupyter notebooks documenting advanced feature engineering and model evaluation steps for both the Titanic and Telco Churn data sets.

Learning Resources

- Consult the recently assigned readings and sample notebooks regarding these data sets.
- Consult relevant documentation, including:
  - GridSearchCV
    https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html
  - Scikit Learn Metrics and Scoring
    https://scikit-learn.org/stable/modules/model_evaluation.html
  - Etc.
- Consult other resources as desired, including others' Kaggle notebooks, etc.

Eminent Readability

Your final notebooks should be showpieces that can be read, understood, and appreciated by data analysts regardless of their Python knowledge. Thus, throughout the notebooks:

1. Provide well-structured headings to guide the reader.
2. Add well written markdown to communicate your actions, expectations, interpretations, evaluations, and decisions.

BONUS Options

Specific requirements for each data set are below. But beyond those, if time and energy allow, you are encouraged to further enhance your notebooks to make them even *showier* show-pieces.

For example, you might:

- Improve code and/or code output.
- Pull validation and test performance metrics into a dataframe for sorting and evaluation.
- Visualize performance with plots to aid comparison and interpretation.
- Etc.

Titanic Notebook Requirements and Options

Combine data preparation and modeling steps into a single notebook. Enhance the previously completed preparation and modeling steps in the following ways:

Data Preparation

1. Create binned versions of Age and Fare, with somewhere between 5 to 15 bins each. Explain how and why your strategy should contribute to better model training. Use appropriate plots to illustrate.
2. Create fewer groupings for Family_Count. Explain how and why your strategy should contribute to better model training. Use appropriate plots to illustrate.
3. Create a new Titles feature, extracted from Names. Group titles into approximately 4-5 categories. Explain how and why your strategy should contribute to better model training. Use appropriate plots to illustrate.
4. OPTIONAL: Create a new Is_Married feature.
5. OPTIONAL: Create a new Deck feature by extracting deck from Cabin.
6. OPTIONAL: Use SMOTE to balance target variable classes.

Category to Numeric Conversion

- Use One-Hot or Label encoding, as appropriate to the data.

Modeling

1. Use at least these algorithms:
   a. Logistic Regression
   b. Random Forest
   c. Multi-Layer Perceptron
   d. Gradient Boosting
2. OPTIONAL: Include additional algorithms, such as:
   a. XGBoost
   b. LightGBM

Model Evaluation

1. In training, validation, and testing, continue to prioritize ***Accuracy***, as this is the metric typically prioritized with this data set.
2. Add the following model evaluation metrics and plots to the validation and test process:
   a. Confusion Matrix
   b. ROC Curve
   c. AUC Score
   d. F1 Score
3. When two models have similar performance for Accuracy, interpret the implications of the other scores.

Telco Churn Notebook Requirements and Options

Enhance the previously completed preparation and modeling steps in the following ways:

Data Preparation

1. Evaluate others' strategies and choose 2-3 feature engineering steps that strike you as potentially most promising. These steps should include:
    a. Construct new, improved features from current features.
    b. Remove features that correlate too strongly with others (reducing multicollinearity).
    c. Etc.
2. For all new steps, provide appropriate plots and discussion to clarify the expected contribution to enhanced model performance.
3. Use SMOTE to balance target variable classes.

Category to Numeric Conversion

- Use One-Hot or Label encoding, as appropriate to the data.

Modeling

1. Use at least these algorithms:
    a. Logistic Regression
    b. Random Forest
    c. Multi-Layer Perceptron
    d. Gradient Boosting
2. OPTIONAL: Include additional algorithms, such as:
    a. XGBoost
    b. LightGBM

Model Evaluation

1. In training, validation, and testing, prioritize the best balance of Accuracy, Precision, and Recall — *with special emphasis on **recall*** — as our business goal is to reach out to as many potential churners as possible.
2. During training, adjust GridSearchCV to prioritize the AUC score. (OPTIONAL: Add an additional step to compare model results when the F1 score is used.)
3. Add the following model evaluation metrics and plots to the validation and test process:
    a. Confusion Matrix
    b. ROC Curve
    c. AUC Score
    d. F1 Score
4. During validation and testing, seek the model with (a) the best AUC (or F1) score, and (b) the best recall score.

What to Submit

Submit your two completed notebooks:

1. Titanic Final
2. Telco Churn Final

If convenient, also include HTML exports of your notebooks.