CLICK-THROUGH RATE PREDICTION AND ANALYSIS

CALTECH-DATA ANALYTICS CERTIFICATION PROGRAM

-AYESHA SIDDIQUA

CONTENTS

- Objective and Problem Statement
- Model building
- Classification Methods
- Model Evaluation
- Importance of errors made in prediction of the model
- Conclusion

OBJECTIVE AND PROBLEM STATEMENT

- Most of the websites we visit include ads. The online advertising industry is huge, and players such as Google, Amazon, and Facebook generate billions of dollars by targeting the correct audiences with relevant ads. Most of the decisions about ads are data-driven solutions.
- As part of this assignment, we are required to predict whether a user will click on an ad or not using classification models as they are best suited here.
- An important exercise marketing companies need to do before making any of the above decisions is a click-through rate (CTR) prediction. The objective is to predict whether the audience will click on an ad or not and thus help the marketing team answer ad placement-related questions.

MODEL BUILDING AND CLASSIFICATION METHODS IMPLEMENTED:

Model Building: Initially the data was divided in 3 parts. For training, validating and testing the model in 80-10-10 ratios respectively. To improve the model performance, the date=a was divided in 70-30 ration for training and testing (In decision tree classifier)

Classification methods:

- Logistic regression
- Decision Tree classifier
- Random Forest classifier

INTERPRET THE EVALUATION METRIC OF CHOICE

- In the decision tree classifier model, we have used the F1 score as the evaluation metric of choice.
- The F1 score is a measure of the balance between precision and recall, and provides a single number that summarizes the overall performance of the model. In this case, the F1 score of 0.74 indicates that the model is making both false positive and false negative errors, although it is striking a reasonable balance between the two. This means that the model is correctly identifying some instances of the target class (precision) while also minimizing the number of missed instances (recall).
- However, the F1 score alone may not provide a complete picture of the model's performance, and it is important to consider other evaluation metrics as well. (This has been mentioned in the upcoming slides)

IMPORTANCE OF ERRORS MADE IN PREDICTION OF THE MODEL

- The implications of errors made by a model in prediction depend on the specific application and the type of errors made. In the case of the above decision tree classifier, the evaluation metrics (accuracy, precision, recall, F1 score, and confusion matrix) can provide some insights into the type and severity of errors made by the model.

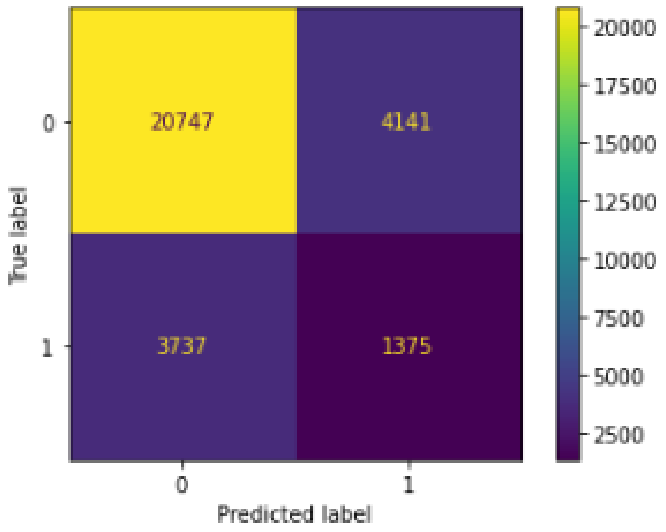
Here's a brief explanation of the implications of each evaluation metric:

- Accuracy: This metric measures the overall accuracy of the model in predicting the correct class labels. A high accuracy score (close to 1.0) indicates that the model is performing well, while a low accuracy score (close to 0.0) indicates that the model is making a large number of errors.
- Precision: This metric measures the proportion of true positive predictions among all positive predictions made by the model. A high precision score (close to 1.0) indicates that the model is making very few false positive errors, while a low precision score (close to 0.0) indicates that the model is making a large number of false positive errors.

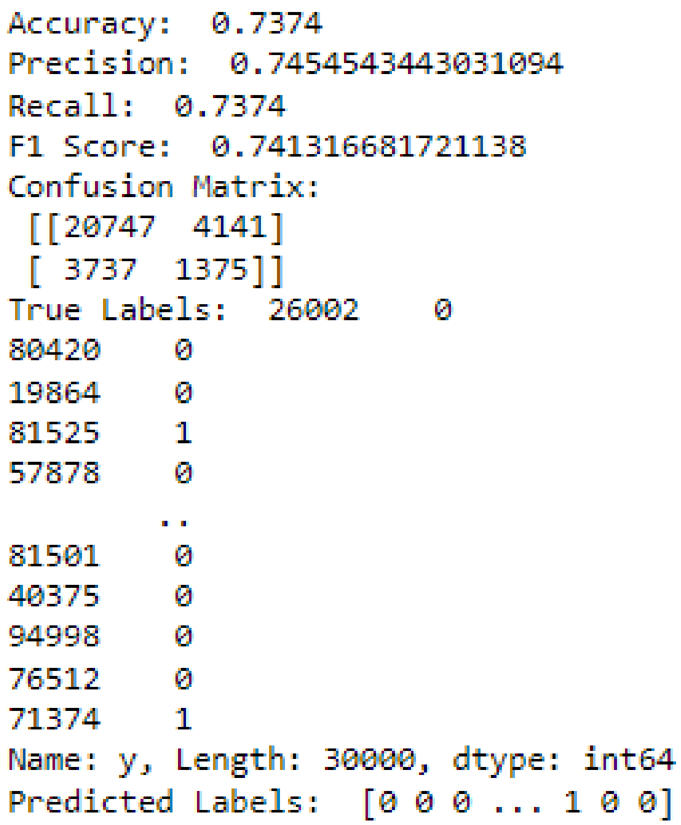
IMPORTANCE OF ERRORS MADE IN PREDICTION OF THE MODEL

- Recall: This metric measures the proportion of true positive predictions among all actual positive instances in the dataset. A high recall score (close to 1.0) indicates that the model is making very few false negative errors, while a low recall score (close to 0.0) indicates that the model is missing a large number of actual positive instances.
- F1 score: This is the harmonic mean of precision and recall, and provides a balance between the two metrics. A high F1 score (close to 1.0) indicates that the model is making both very few false positive and false negative errors, while a low F1 score (close to 0.0) indicates that the model is making a large number of either type of error.
- Confusion matrix: This matrix provides a more detailed breakdown of the number of true positive, false positive, true negative, and false negative predictions made by the model. This can help to identify which classes are being confused by the model and which types of errors are being made.

EVALUATING THE MODEL SELECTED-DECISION TREE CLASSIFER



Decision tree classifier performs fairly better than other classification models.



```
Accuracy:  0.7374
Precision:  0.7454543443031094
Recall:  0.7374
F1 Score:  0.741316681721138
Confusion Matrix:
 [[20747  4141]
 [ 3737  1375]]
True Labels:  26002    0
80420    0
19864    0
81525    1
57878    0
        ..
81501    0
40375    0
94998    0
76512    0
71374    1
Name: y, Length: 30000, dtype: int64
Predicted Labels:  [0 0 0 ... 1 0 0]
```

CONCLUSION:

- The data given consisted several numeric and categorical attributes. Data cleaning was performed and after stage 1 EDA, data was analyzed to remove unnecessary columns
- Categorical data was converted to numeric using hash function.
- Dataset was divided into training, validation and testing data.
- After model building, 3 classification/ prediction models were implemented namely- Logistic Regression, Decision Tree Classifier and Random Forest Classifier.
- Random Forest classifier and Decision tree classifier were selected for further steps to improve the model efficiency.
- Upon cross validation and calculating evaluation metrics for both, Decision Tree classifier was observed to be a better model. The model performances do not seem to be that good as the data was extremely biased and unbalanced.