Sri Lanka Institute of Information Technology

Transaction Cancellation Prediction

Artificial Intelligence and Machine Learning| IT2011

Year 2 semester 1 (2025)

Group ID :- **2025-Y2-S1-MLB-B11G1-06**

Group Members

Jayasinghe T T A	IT24102807
Nayakarathne R.M.U.K	IT24102939
Nayakarathna N.M.S.S	IT24102925
Rasanjana S.P	IT24102936
Maheepala S A D A O	IT24102901
Sivasankar P	IT24102973

Table of Contents

Member 1: IT24102901 – Decision Tree and Random Forest models

First of all, after the pre-processing was done on the dataset, another 1000 records were separated for further testing purposes.

Variant 1:

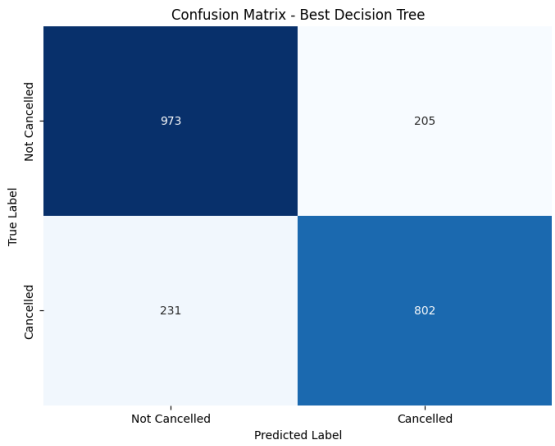
The best fit combination of the hyper-parameter for the decision tree was found from the following parameter grid.

'criterion': ['gini', 'entropy']

'max_depth': [5, 10, 15, None]

'min_samples_split': [2, 5, 10]

GridSearch CV was used for validation with 5 folds for each 24 candidates, totaling 120 fits. The best combination was found to be: 'criterion': 'gini', 'max_depth': 15, 'min_samples_split': 2.




```
Evaluating the best Decision Tree model on the test set...
Accuracy: 0.8028
Precision: 0.7964
Recall: 0.7764
F1-Score: 0.7863
ROC AUC Score: 0.8019

Confusion Matrix:
[[973 205]
 [231 802]]

Classification Report:
              precision    recall  f1-score   support

       False       0.81      0.83      0.82      1178
        True       0.80      0.78      0.79      1033

    accuracy                           0.80      2211
   macro avg       0.80      0.80      0.80      2211
weighted avg       0.80      0.80      0.80      2211
```

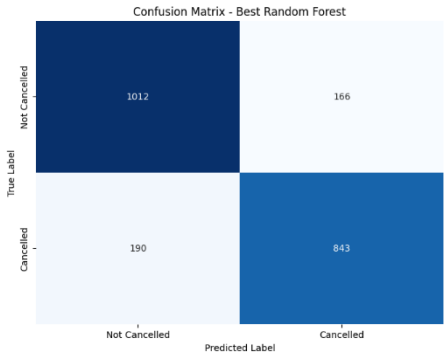
The same was done for the Random Forest model as well. GridSearch CV was used for validation with 5 folds for each 12 candidate, totaling 60 fits.

'n_estimators': [100, 200]

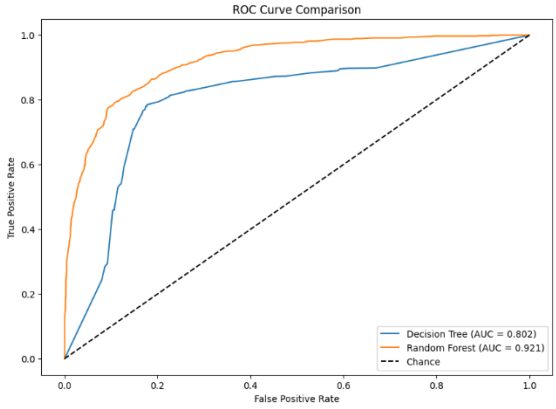
'max_depth': [10, 20, None]

'min_samples_split': [2, 5]

The best combination was found to be: 'max_depth': None, 'min_samples_split': 2, 'n_estimators': 200.




```
Evaluating the best Random Forest model on the test set...
Accuracy: 0.8390
Precision: 0.8355
Recall: 0.8161
F1-Score: 0.8257
ROC AUC Score: 0.9210

Confusion Matrix:
[[1012  166]
 [ 190  843]]

Classification Report:
              precision    recall  f1-score   support

       False       0.84      0.86      0.85      1178
        True       0.84      0.82      0.83      1033

    accuracy                           0.84      2211
   macro avg       0.84      0.84      0.84      2211
weighted avg       0.84      0.84      0.84      2211
```



Finally, the models were tested on the unseen data which was separated earlier. Through the evaluation metrics it was visible that the models were overfitted as they did not generalize well.

```
Evaluating the best Decision Tree model on the unseen test set...

--- Decision Tree Classification Report (Unseen Test Set) ---
              precision    recall  f1-score   support

Not Cancelled       0.82      0.82      0.82       533
    Cancelled       0.79      0.80      0.80       467

     accuracy                           0.81      1000
    macro avg       0.81      0.81      0.81      1000
 weighted avg       0.81      0.81      0.81      1000

Decision Tree F1-Score (Unseen Test Set): 0.7962

Evaluating the best Random Forest model on the unseen test set...

--- Random Forest Classification Report (Unseen Test Set) ---
              precision    recall  f1-score   support

Not Cancelled       0.85      0.86      0.86       533
    Cancelled       0.84      0.83      0.83       467

     accuracy                           0.85      1000
    macro avg       0.85      0.84      0.85      1000
 weighted avg       0.85      0.85      0.85      1000

Random Forest F1-Score (Unseen Test Set): 0.8341
```

--- F1 Score Comparison (Training vs. Unseen Test Set) ---

	Model	F1-Score (Training)	F1-Score (Unseen Test Set)
0	Decision Tree	0.927782	0.796158
1	Random Forest	0.999879	0.834052

Variant 2:

Since the model was found out to be overfitted previously, necessary precautions were taken to mitigate them.

There is high dimensionality in the encoded TF-IDF columns, therefore dimensionality reduction has been done using PCA to shrink many TF-IDF text columns down to just 50 components. This captures the most important patterns from the text data in a much smaller, simpler format.

In addition, GridSearchCV was used here as well for further cross-validation and to stop over-fitting.

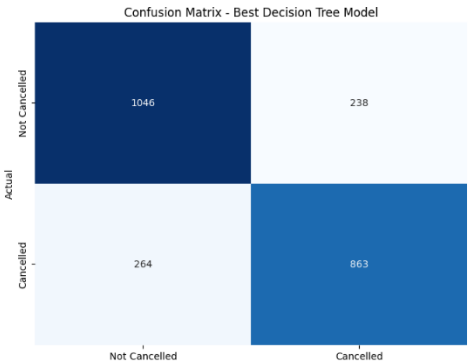
Furthermore, model complexity was reduced as the previous hyperparameters allowed the tree to grow thereby allowing it to memorize data.

```
# Parameter grid for Decision Tree
param_grid_dt = {
    'criterion': ['gini', 'entropy'],
    'max_depth': [10, 15, 20],
    'min_samples_split': [2, 5, 10]
}
```

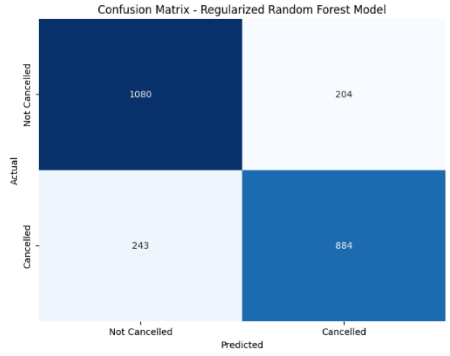
```
# Parameter grid for Random Forest
param_grid_rf_regularized = {
    'n_estimators': [100, 200],
    'max_depth': [5, 8, 10],
    'min_samples_split': [10, 20, 40]
}
```

Best parameters for Regularized Random Forest:
{'max_depth': 10, 'min_samples_split': 10, 'n_estimators': 200}

```
Best parameters for Decision Tree: {'criterion': 'gini', 'max_depth': 10, 'min_samples_split': 10}
Best ROC AUC score on training data (Decision Tree): 0.8270337583566102

--- Best Decision Tree Model Evaluation ---
Classification Report:
              precision    recall  f1-score   support

       False       0.80      0.81      0.81      1284
        True       0.78      0.77      0.77      1127

    accuracy                           0.79      2411
   macro avg       0.79      0.79      0.79      2411
weighted avg       0.79      0.79      0.79      2411

Confusion Matrix:
[[1046  238]
 [ 264  863]]
ROC AUC Score: 0.8471485099525387
Average Precision Score: 0.7725853916581836
Accuracy Score: 0.7917876399834094
```




```
Best parameters for Regularized Random Forest: {'max_depth': 10, 'min_samples_split': 10, 'n_estimators': 200}
Best ROC AUC score on training data: 0.8800516833891985

--- Regularized Random Forest Model Evaluation on Test Set ---
Classification Report:
              precision    recall  f1-score   support

       False       0.82      0.84      0.83      1284
        True       0.81      0.78      0.80      1127

    accuracy                           0.81      2411
   macro avg       0.81      0.81      0.81      2411
weighted avg       0.81      0.81      0.81      2411

Confusion Matrix:
[[1080  204]
 [ 243  884]]
ROC AUC Score: 0.8962785439246808
Average Precision Score: 0.881848081936408
Accuracy Score: 0.8145997511406056
```

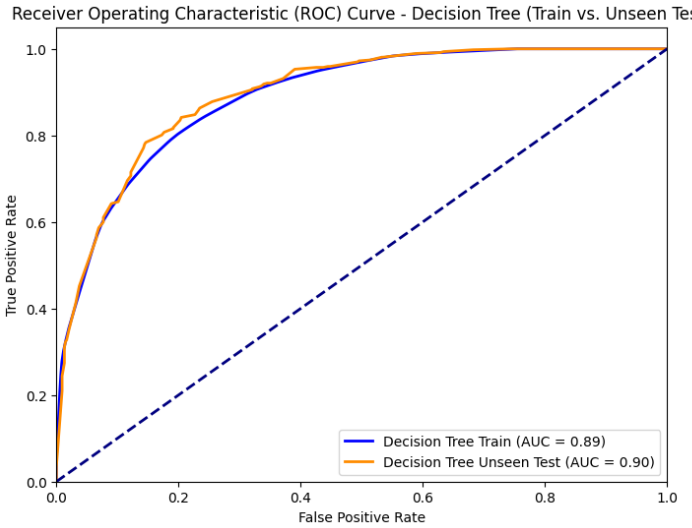


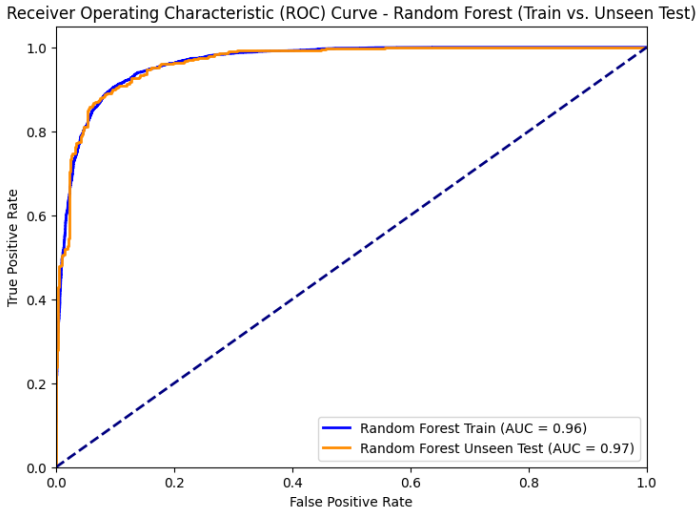
Later through manual tunning it was found that the best parameters for Regularized Decision Tree: 'criterion': 'entropy', 'max_depth': 8, 'min_samples_split': 60.

```
--- Regularized Decision Tree Model Evaluation on Test Set ---
Classification Report:
              precision    recall  f1-score   support

       False       0.79      0.79      0.79      1284
        True       0.76      0.76      0.76      1127

    accuracy                           0.78      2411
   macro avg       0.78      0.77      0.77      2411
weighted avg       0.78      0.78      0.78      2411

Confusion Matrix:
[[1020  264]
 [ 276  851]]
ROC AUC Score: 0.8558847960151147
Average Precision Score: 0.8177324606160715
Accuracy Score: 0.7760265450020738
```

This variant had the best score for the metrics so far with proper generalization and no over-fitting. The two models were tested against the separated unseen dataset and yielded the following results.

```
--- Tuned Decision Tree Model Evaluation on Unseen Test Set ---
Classification Report:
              precision    recall  f1-score   support

       False       0.83      0.82      0.83       533
        True       0.80      0.81      0.80       467

    accuracy                           0.82      1000
   macro avg       0.82      0.82      0.82      1000
weighted avg       0.82      0.82      0.82      1000

F1 Score on Unseen Test Data (Decision Tree): 0.8038
ROC AUC Score on Unseen Test Data (Decision Tree): 0.8962

--- Regularized Random Forest Model Evaluation on Unseen Test Set ---
Classification Report:
              precision    recall  f1-score   support

       False       0.90      0.91      0.91       533
        True       0.90      0.89      0.89       467

    accuracy                           0.90      1000
   macro avg       0.90      0.90      0.90      1000
weighted avg       0.90      0.90      0.90      1000

F1 Score on Unseen Test Data (Random Forest): 0.8937
ROC AUC Score on Unseen Test Data (Random Forest): 0.9651
```






```
--- Performance Comparison (Train vs. Unseen Test) ---
Decision Tree - F1 Score (Training): 0.8700
Decision Tree - F1 Score (Unseen Test): 0.8038
Decision Tree - ROC AUC Score (Training): 0.9400
Decision Tree - ROC AUC Score (Unseen Test): 0.8962

Random Forest - F1 Score (Training): 0.9000
Random Forest - F1 Score (Unseen Test): 0.8937
Random Forest - ROC AUC Score (Training): 0.9700
Random Forest - ROC AUC Score (Unseen Test): 0.9651
```

Variant 3:

For the third variant, the dataset version including StockCode-based features was utilized, which introduced additional encoded variables and consequently higher dimensionality.

As with the previous iteration, PCA was applied to reduce the large number of TF-IDF text features to 50 principal components, retaining the most informative variance while simplifying the feature space.

GridSearchCV was again employed for hyperparameter optimization and to maintain robust cross-validation performance. The model's complexity was carefully controlled to prevent overfitting through the use of regularization and parameter constraints.

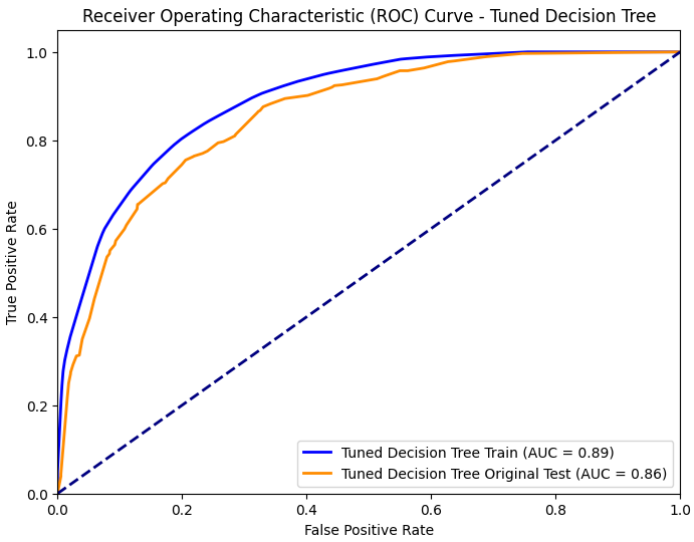
The best hyper-parameters were found to be the same values as in the variant 2. Since they yielded good performance results they were retained.

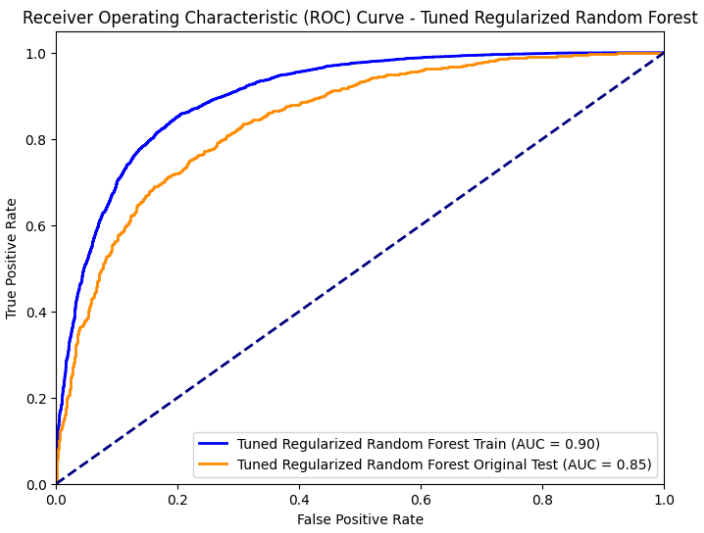
However, despite these optimizations, Variant 3 showed slightly inferior generalization performance compared to Variant 2. T

he inclusion of StockCode features likely added redundant or weakly correlated information, introducing mild noise into the learning process.

As a result, while the model still demonstrated solid predictive capability, its ROC AUC scores on unseen data were slightly lower, indicating that Variant 2's feature representation was cleaner and more efficient for capturing the underlying patterns without unnecessary complexity.

```
--- Performance Comparison (Training vs. Original Test) ---

Decision Tree:
  F1 Score (Training): 0.7885
  ROC AUC Score (Training): 0.8900
  F1 Score (Original Test): 0.7591
  ROC AUC Score (Original Test): 0.8564

Random Forest:
  F1 Score (Training): 0.7771
  ROC AUC Score (Training): 0.9017
  F1 Score (Original Test): 0.7070
  ROC AUC Score (Original Test): 0.8457
```





In conclusion, the variant 2 of the random forest model is the optimum model out of all the other variations.

Best Regularized Random Forest: 'max_depth': 10, 'min_samples_split': 10, 'n_estimators': 200.