SpeakX Project Report

Data Preprocessing:

Data Preprocessing includes:

1. Removing CustomerID from feature list.
2. Check for missing values and update them.
3. Converting categorical variables into numerical values using techniques like One-Hot Encoding.
4. Deletion of entries having '0' tenure duration.

EDA:

EDA found:

1. 'No' for Churn value is far more common (73%)
2. This percentage was almost the same for both male and female customers.
3. Those having short-term contracts, i.e. month-to-month contracts are far more likely to Churn followed by two-year and one-year contracts.
4. Customers using Electronic check are more likely to Churn as compared to those getting mailed check, or using credit card or bank transfer.
5. When it comes to Internet service, those having Fibre optic service are more likely to Churn, followed by those using DSL, followed by those with No internet service, all having significant differences from each other.
6. Those not having dependents are more likely to Churn.
7. Those not having partners are more likely to Churn. So the order of likelihood of Churn becomes: Those having dependents > Those having partners > Those not having partners (single users).
8. Senior citizens are more likely to Churn.
9. Those opting for paperless billing are more likely to Churn.
10. Those not opting for Tech Support are more likely to Churn compared to those opting for it followed by those with no internet service.

11. Whether the customer is opting for Phone Service or not does not make a big difference in the likelihood of Churn.
12. Customers in the first 15 tenure values are more likely to Churn and the last 15 tenure values are less likely to Churn. So, Likelihood of Churn decreases with increasing tenure values.

Churn Prediction:

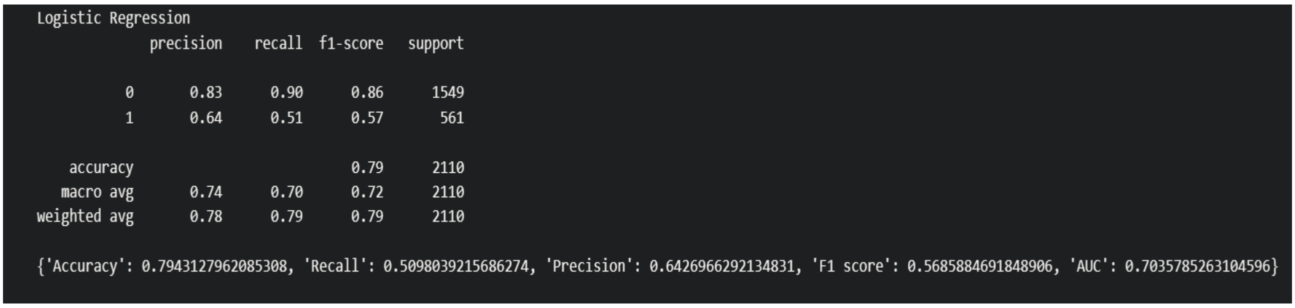
Churn prediction comes under classification task, so a few classification algorithms are under consideration**:**

1. Linear Regression
2. Decision Tree Classification
3. Random Forest Classification
4. Support Vector Machine
5. XGBoost Classifier

Logistic Regression:

max_iter=1000
penalty: L2

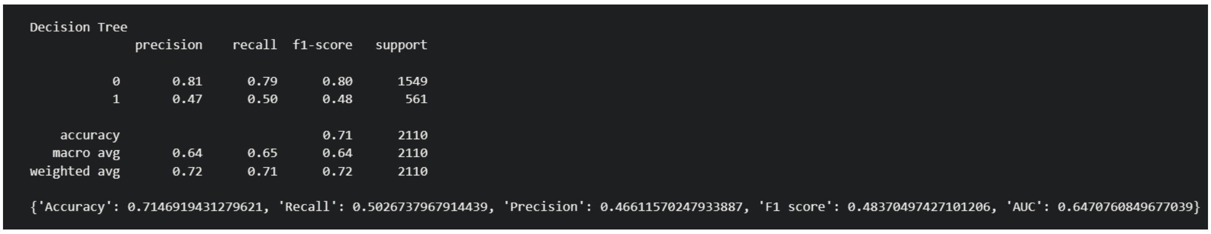
Results:



```
Logistic Regression
              precision    recall  f1-score   support

           0       0.83      0.90      0.86      1549
           1       0.64      0.51      0.57       561

    accuracy                           0.79      2110
   macro avg       0.74      0.70      0.72      2110
weighted avg       0.78      0.79      0.79      2110

{'Accuracy': 0.7943127962085308, 'Recall': 0.5098039215686274, 'Precision': 0.6426966292134831, 'F1 score': 0.5685884691848906, 'AUC': 0.7035785263104596}
```

Decision Tree Classification:
Criterion: ‘gini’
max_depth: None
splitter: Best

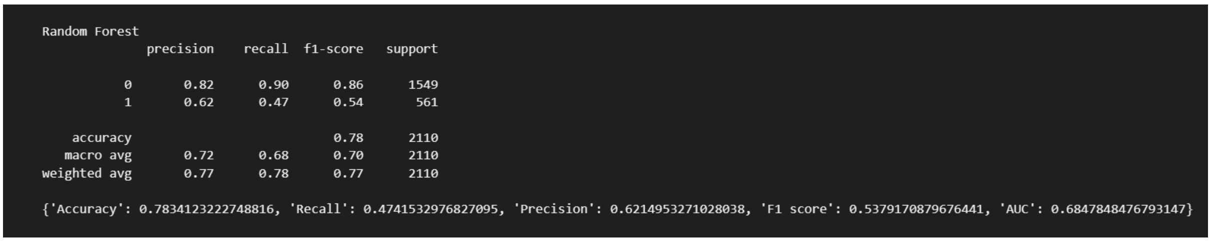
Results:



```
Decision Tree
              precision    recall  f1-score   support

           0       0.81      0.79      0.80      1549
           1       0.47      0.50      0.48       561

    accuracy                           0.71      2110
   macro avg       0.64      0.65      0.64      2110
weighted avg       0.72      0.71      0.72      2110

{'Accuracy': 0.7146919431279621, 'Recall': 0.5026737967914439, 'Precision': 0.46611570247933887, 'F1 score': 0.48370497427101206, 'AUC': 0.6470760849677039}
```

Random Forest Classification:
n_estimators = 100
criterion = ‘gini’

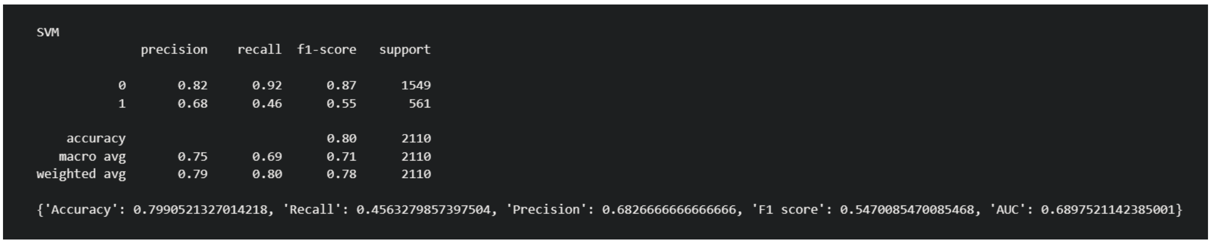
Results:



```
Random Forest
              precision    recall  f1-score   support

           0       0.82      0.90      0.86      1549
           1       0.62      0.47      0.54       561

    accuracy                           0.78      2110
   macro avg       0.72      0.68      0.70      2110
weighted avg       0.77      0.78      0.77      2110

{'Accuracy': 0.7834123222748816, 'Recall': 0.4741532976827095, 'Precision': 0.6214953271028038, 'F1 score': 0.5379170879676441, 'AUC': 0.6847848476793147}
```

Support Vector Machine:
C = 1
kernel: Linear
gamma: scale

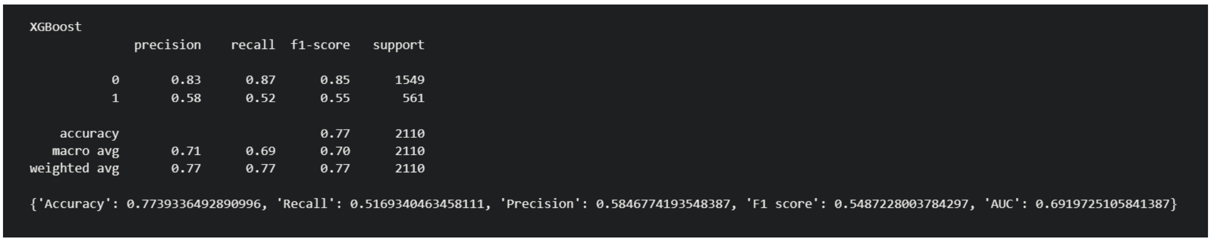
Results:



```
SVM
              precision    recall  f1-score   support

           0       0.82      0.92      0.87      1549
           1       0.68      0.46      0.55       561

    accuracy                           0.80      2110
   macro avg       0.75      0.69      0.71      2110
weighted avg       0.79      0.80      0.78      2110

{'Accuracy': 0.7990521327014218, 'Recall': 0.4563279857397504, 'Precision': 0.6826666666666666, 'F1 score': 0.5470085470085468, 'AUC': 0.6897521142385001}
```

XGBoost Classification:
booster=‘gbtree’
colsample_bylevel=1
colsample_bynode=1
colsample_bytree=1

Results:



```
XGBoost
              precision    recall  f1-score   support

           0       0.83      0.87      0.85      1549
           1       0.58      0.52      0.55       561

    accuracy                           0.77      2110
   macro avg       0.71      0.69      0.70      2110
weighted avg       0.77      0.77      0.77      2110

{'Accuracy': 0.7739336492890996, 'Recall': 0.5169340463458111, 'Precision': 0.5846774193548387, 'F1 score': 0.5487228003784297, 'AUC': 0.6919725105841387}
```

★ Best F1 score: 0.5688 (Logistic Regression)

Post-prediction Feature Analysis:

Findings:

1. Logistic Regression: Monthly Charges, tenure and Contract are the most important features.
2. Decision Tree Classifier: Monthly Charges, Total Charges, tenure and Contract are the most important features.
3. Random Forest: Tenure, Monthly Charges and Total Charges are the most important features.
4. SVM: Monthly Charges, tenure, Internet Service are the most important features.
5. XGBoost Classifier: Contract is by far the most important feature.

Here, by ‘important’ features, we mean features with the most amount of information.