CS 613 Machine Learning Project: Predicting Churn of Telecommunications Customers

By -
Shubham Jadhav (sj3237)
Margaret Brier (mgb87)
Yash Gupta (yg444)

Problem

Churn is a one of the biggest problem for companies that provide subscription based services especially in the telecom industry.
Research has shown that the average monthly churn rate among the top 4 wireless carriers in the US is 1.9% - 2%.
The Churn Prediction modeling project aims to identify customers likely to cancel their subscriptions or stop using a service.
Utilizing a dataset with various customer attributes and their churn status, this project employs machine learning techniques to predict churn probability.
The insights gained from the analysis will enable the development of targeted strategies to improve customer retention and reduce turnover rates.

Prior Work

Enhanced Logistic Regression Techniques: Arivazhagan et al. (2020) showed that logistic regression enhanced with regularization and optimization is effective for churn prediction, highlighting its adaptability for binary outcomes in telecommunications.

Decision Trees for Churn Analysis: Decision trees are praised for their ability to map decision paths intuitively, aiding in the identification of key factors influencing churn, and providing a simple yet interpretable tool for understanding customer behavior.

Advances in Ensemble Methods: Ensemble methods like random forests improve churn prediction by combining multiple decision tree models to reduce variance and bias, demonstrating the power of these methods in addressing complex customer data relationships.

Importance of Feature Selection and Engineering: The selection of crucial predictors such as service tenure and contract type, along with advanced data preprocessing, is key to optimizing churn prediction models, emphasizing the importance of detailed feature analysis.

Challenges in Model Adaptation: Adapting models to the changing patterns of customer behavior and preferences remains a significant challenge, necessitating ongoing refinement to maintain accuracy and relevance in churn prediction.

Basic Approach

Multiple models were used to predict the probability of customer churn in the range {0,1}, where Churn = 1 means the customer left and Churn = 0 means the customer has not left.

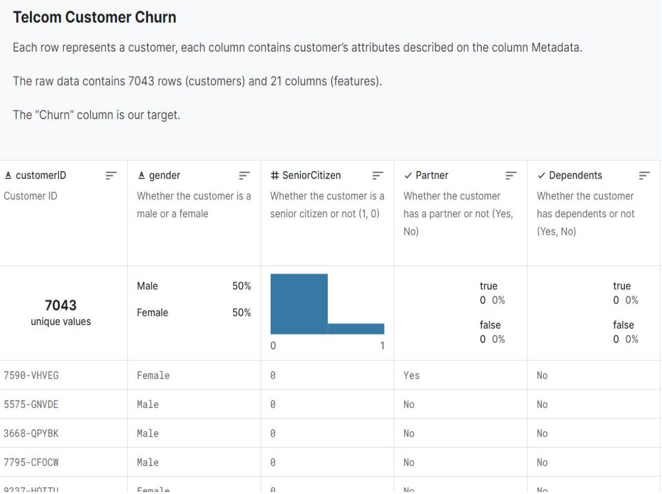
Logistic Regression and Decision Tree algorithms were applied to the dataset.

A Random Forest ensemble was constructed using Decision Trees.

An Ensemble model was created using the logistic regression and random forest models.

DataSet

https://www.kaggle.com/datasets/blastchar/telco-customer-churn



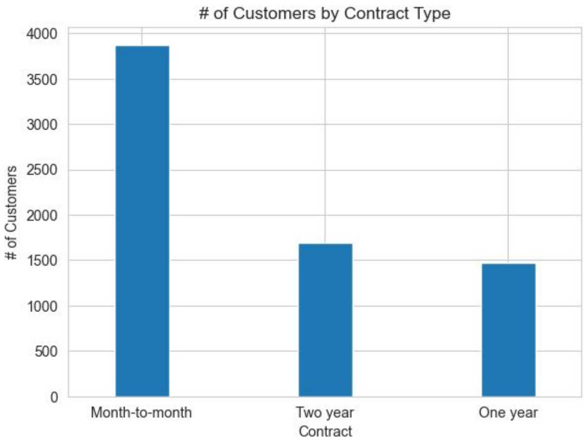
Telcom Customer Churn

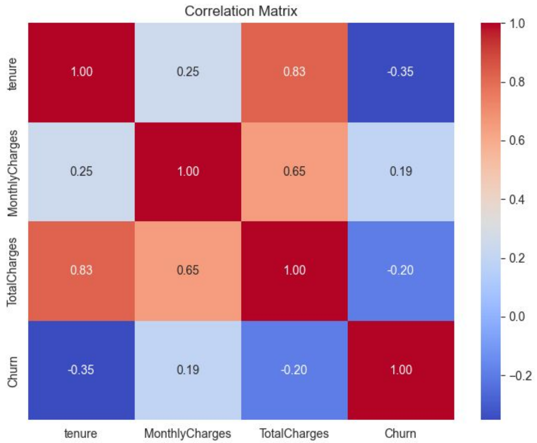
Each row represents a customer, each column contains customer's attributes described on the column Metadata.

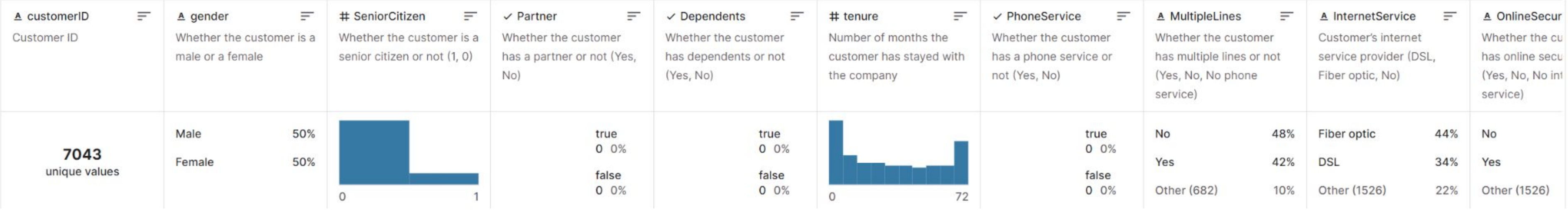
The raw data contains 7043 rows (customers) and 21 columns (features).

The "Churn" column is our target.

A customerID	A gender	# SeniorCitizen	✓ Partner	✓ Dependents
Customer ID	Whether the customer is a male or a female	Whether the customer is a senior citizen or not (1, 0)	Whether the customer has a partner or not (Yes, No)	Whether the customer has dependents or not (Yes, No)
7043 unique values	Male 50% Female 50%	0 1	true 0 0% false 0 0%	true 0 0% false 0 0%
7590-VHVEG	Female	0	Yes	No
5575-GNVDE	Male	0	No	No
3668-QPYBK	Male	0	No	No
7795-CFOCW	Male	0	No	No
9237-HQITU	Female	0	No	No

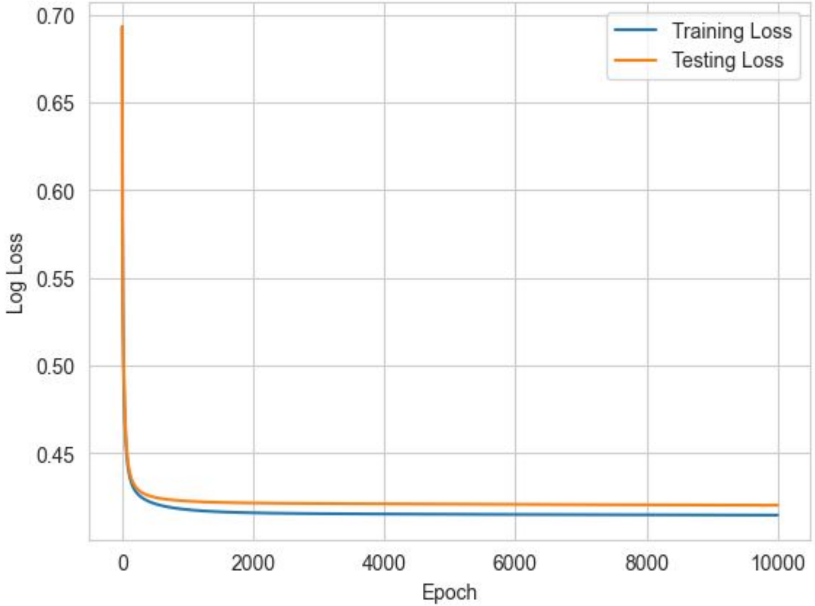






A customerID	A gender	# SeniorCitizen	✓ Partner	✓ Dependents	# tenure	✓ PhoneService	A MultipleLines	A InternetService	A OnlineSecur
Customer ID	Whether the customer is a male or a female	Whether the customer is a senior citizen or not (1, 0)	Whether the customer has a partner or not (Yes, No)	Whether the customer has dependents or not (Yes, No)	Number of months the customer has stayed with the company	Whether the customer has a phone service or not (Yes, No)	Whether the customer has multiple lines or not (Yes, No, No phone service)	Customer's internet service provider (DSL, Fiber optic, No)	Whether the cu has online secu (Yes, No, No int service)
7043 unique values	Male 50% Female 50%	0 1	true 0 0% false 0 0%	true 0 0% false 0 0%	0 72	true 0 0% false 0 0%	No 48% Yes 42% Other (682) 10%	Fiber optic 44% DSL 34% Other (1526) 22%	No Yes Other (1526)

Logistic Regression Results



Logistic Regression model accuracy: 0.8041

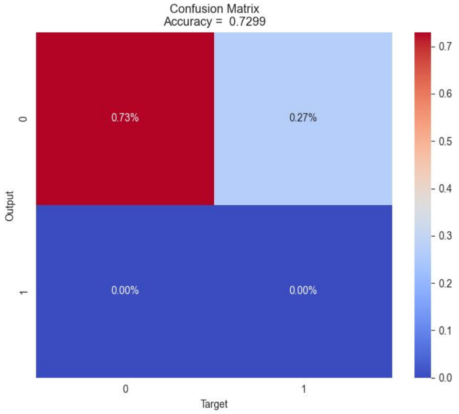
Precision: 0.6719

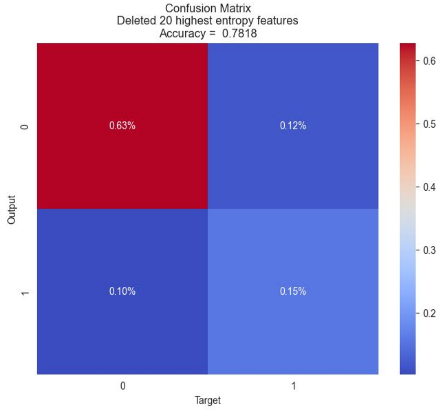
Recall: 0.5273

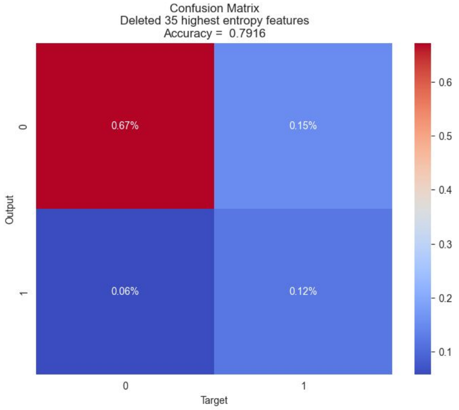
F-Measure: 0.5909

Logistic Regression demonstrated the highest overall accuracy and precision, indicating its effectiveness in correctly identifying churn and non-churn customers with a minimal number of false positives.

Decision Tree Results


Confusion Matrix
Accuracy = 0.7299


Confusion Matrix
Deleted 20 highest entropy features
Accuracy = 0.7818


Confusion Matrix
Deleted 35 highest entropy features
Accuracy = 0.7916

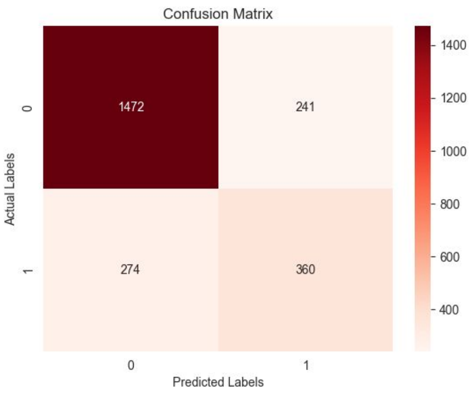
All 41 features
Overall accuracy of validation data is 0.7298
Confusion Matrix values:
[1713 634]
 [0   0]]

20 lowest-entropy features deleted
21 features remain
Overall accuracy of validation data is 0.7819
Confusion Matrix values:
[[1473 272]
 [240 362]]

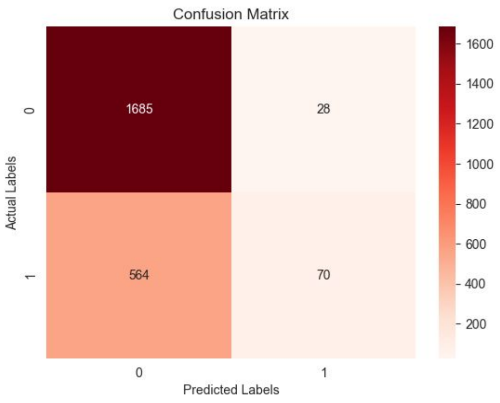
35 lowest-entropy features deleted
6 features remain
Overall accuracy of validation data is 0.7916
Confusion Matrix values:
[[1576 352]
 [137 282]]

The Decision Tree model showed lower performance compared to Logistic Regression but revealed that the model could be improved with feature reduction.

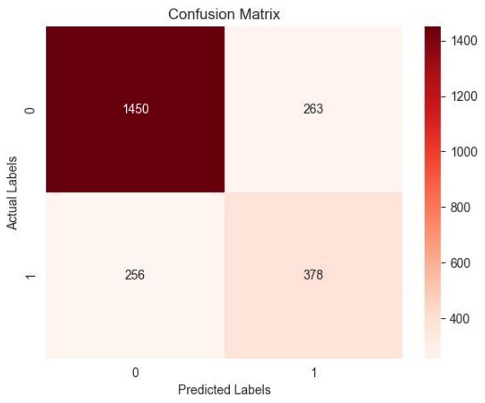
Random Forest Results



```
There are 41 features
Trees: 500, Max Depth: 41,
Features: 41
Training
Accuracy: 0.7527683134582623,
Precision: 0.7202380952380952,
Recall: 0.09797570850202429,
F1: 0.17248752672843906
Testing
Accuracy: 0.7477631018321261,
Precision: 0.7142857142857143,
Recall: 0.11041009463722397,
F1: 0.1912568306010929
Accuracy: 74.78%
```

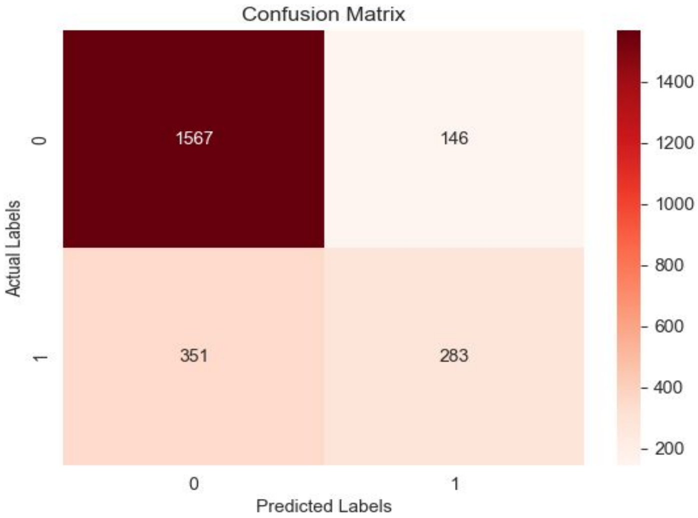


```
There are 41 features
Trees: 100, Max Depth: 4,
Features: 41
Training
Accuracy: 0.770442930153322,
Precision: 0.5652535328345802,
Recall: 0.5506072874493927, F1:
0.5578342904019689
Testing
Accuracy: 0.7788666382616106,
Precision: 0.5897035881435257,
Recall: 0.5962145110410094, F1:
0.5929411764705882
Accuracy: 77.89%
```



```
There are 6 features
Trees: 100, Max Depth: 6,
Features: 6
Training
Accuracy: 0.7617120954003407,
Precision: 0.5379084967320261,
Recall: 0.6663967611336032, F1:
0.5952983725135623
Testing
Accuracy: 0.7682147422241159,
Precision: 0.5556930693069307,
Recall: 0.7082018927444795, F1:
0.622746185852982
Accuracy: 76.82%
```

Random Forest Results



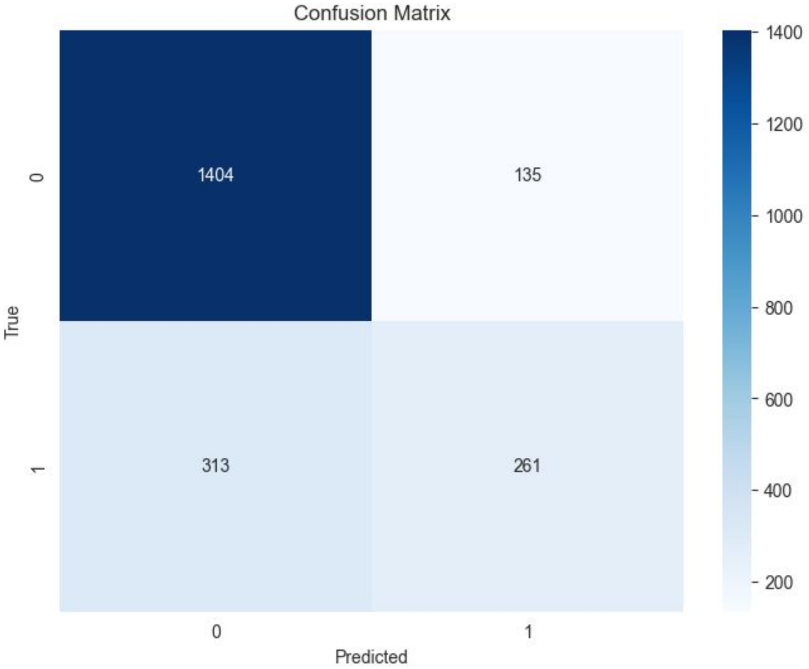
```
There are 21 features, Trees: 500, Max Depth: 21, Features: 21

Training
Accuracy: 0.784923339011925, Precision: 0.6418663303909206, Recall: 0.4121457489878543, F1:
0.5019723865877712

Testing
Accuracy: 0.7882403067746059, Precision: 0.6596736596736597, Recall: 0.44637223974763407, F1:
0.5324553151458137

Accuracy: 78.82%
```

Logistic Regression + Random Forest Ensemble Results



```
Ensemble Model Evaluation:

Accuracy: 0.7880

Precision: 0.6591

Recall: 0.4547

F1 Score: 0.5381
```

Observations

Performance Across Models:

Random Forest showcased the highest recall and F1 Score, demonstrating its superior ability to identify churn customers accurately without excessively misclassifying non-churn customers. This model is adept at handling the complexity of the dataset and avoiding overfitting, which is a common pitfall of Decision Trees.

Logistic Regression had the highest accuracy and showed commendable performance across all metrics. Though, Random Forest had lower accuracy, it performed better overall. This model's strength lies in its simplicity and interpretability, which can be particularly useful when it is important to understand the drivers of churn.

The Decision Tree presented a straightforward model but with lower performance metrics than the Random Forest. It remains valuable for its interpretability and ease of implementation.

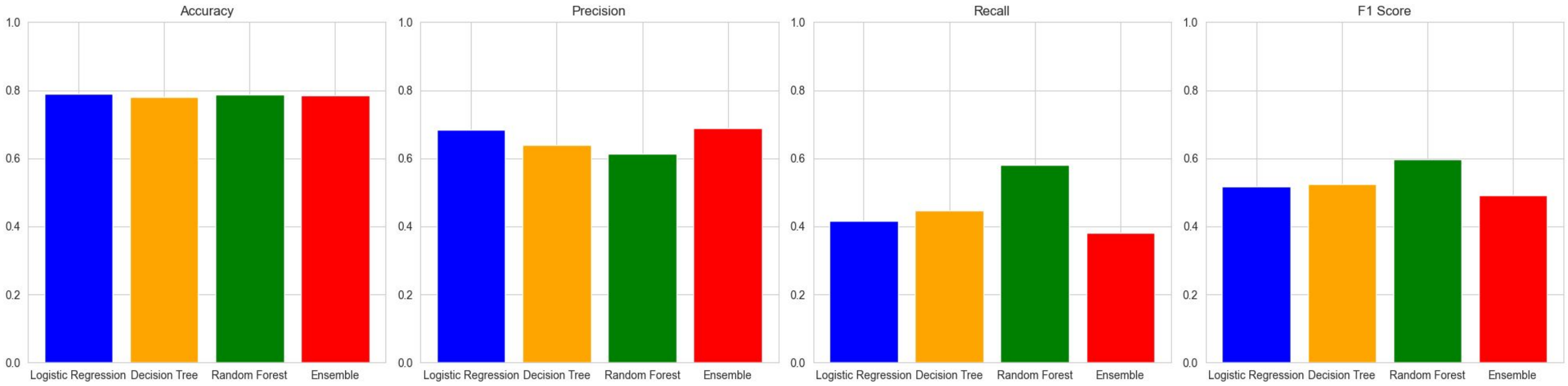
The Ensemble model did not exceed the performance of Random Forest or Logistic Regression. This indicates that the way the models were combined may not have been the most effective. The ensemble method's precision was notably high, but its recall was lacking, suggesting it may not identify as many actual churn cases as Random Forest.

Precision and Recall Trade-off: Despite the high precision of the Ensemble model, its low recall suggests it is less suitable for scenarios where it is critical to capture as many true churn cases as possible. Random Forest, with its balanced recall and precision reflected in the F1 Score, would be preferable in scenarios where both identifying churners and avoiding false churn predictions are important.

F1 Score: Random Forest achieved the highest F1 Score, indicating it maintained a balance between recall and precision. This suggests that Random Forest is the most suitable model for our churn prediction task since it captures the harmonic mean of precision and recall, essential for our dataset.

Results

All models had accuracy ~78-80%



Future Scope

Parameter Tuning: Changing the Number of trees, Max tree depth, Logistic Regression epochs, and Random Forest voting threshold improved model performance, but determining the optimal parameters was not time efficient.

Improved feature reduction: Reducing the number of features in the Decision Tree model based on entropy resulted in higher accuracy. Using feature diversity, feature clustering, or more sophisticated entropy calculations could result in further improvements.

Expansion into Other Markets: Adapting the churn prediction models for use in other subscription-based industries, such as streaming services, to broaden the impact of the research.

Integration of Advanced Machine Learning Models: Exploring the use of more complex models like Gradient Boosting Machines (GBMs) and Deep Learning to improve prediction accuracy.