Prediction of Telecom Customer Churn

Course: MSBA 307- AI for Data Security, Integrity and Risk Mitigation

Final Project

Student: Yichun Chen

Table of Contents

PROBLEM STATEMENT

Customer churn means customer loss; in other words, your customers stop buying your products or cancel subscription to your services. It is a big issue especially in the saturated telecom market where competition is fierce (Zhang et al., 2022). It's essential to preserve customer base because customers are the major component of a business's progress and success (Senthan et al., 2021). When customers switch to alternative service providers, it will lead to decreased profit. Therefore, it has become a critical task for a business to identify potential churners and develop customer retention programs, such as offering promotion coupons and enhancing customer experience by improving customer service.

Customer churn prediction has become a hot research topic in recent years. Churn prediction aims to identify the possible churners in advance before they leave the network. This helps the Customer Relation Management (CRM) department to target these potential churners with effective retention policies so that they will change their mind and stay with the business rather than switching to alternative competitors, therefore reducing the risk of profit loss (Umayaparvathi & Iyakutti, 2016).

Multiple machine learning models have been used to predict customer churn in telecom industry. Senthan et al.(2021) used XGBoost to predict churn in telecommunication industry in Sri Lanka and they obtained an accuracy score of 83.13%. Another study by Dalvi et al. (2016) used logistic regression and decision trees. Random forest model was used by Ullah et al. (2019) with an accuracy score of 88.63%.

This paper aims to predict customer churn of a telecom company, using a dataset published in Kaggle. Two algorithms, random forest and logistic regression, were implemented and compared based on metrics such as recall, f1-score, ROC AUC score, and confusion matrix.

DATASET INFORMATION

Source

The dataset was downloaded from Kaggle (https://www.kaggle.com/datasets/blastchar/telco-customer-churn)

Description

The dataset includes 7043 rows and 21 columns. Each row represents a unique customer, and each column represents the customer's attributes which includes demographic information, service usage information and account information. "Churn" is our target label; there are 15 categorical features and 3 numerical features.

Data Category	Column Name	Description
Demographic:	customerID	unique identification of customer
	gender	whether customer is female or male (female, male)
	SeniorCitizen	Whether the customer is a senior citizen or not (Yes, No)
	Partner	Whether the customer has a partner or not (Yes, No)
	Dependents	Whether the customer has dependents or not (Yes, No)
	tenure	Number of months the customer has stayed with the company
Service Usage:	PhoneService	Whether the customer has a phone service or not (Yes, No)
	MultipleLines	Whether the customer has multiple lines or not (Yes, No, No phone service)
	InternetService	Customer's internet service provider (DSL, Fiber optic, No)
	OnlineSecurity	Whether the customer has online security or not (Yes, No, No internet service)
	OnlineBackup	Whether the customer has online backup or not (Yes, No, No internet service)
	DeviceProtection	Whether the customer has device protection or not (Yes, No, No internet service)
	TechSupport	Whether the customer has tech support or not (Yes, No, No internet service)
	StreamingTV	Whether the customer has streaming TV or not (Yes, No, No internet service)
	StreamingMovies	Whether the customer has streaming movies or not (Yes, No, No internet service)
Account:	Contract	The contract term of the customer (Month-to-month, One year, Two year)

	PaperlessBilling	Whether the customer has signed up paperless billing or not (Yes, No)
	PaymentMethod	The customer's payment method (Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic))
	MonthlyCharges	The amount charged to the customer monthly
	TotalCharges	The total amount charged to the customer
Target Label:	Churn	Whether the customer churned or not (Yes or No)

SOLUTION PROCESS

This project follows the standard data analytics workflow. After data was imported, we used descriptive statistics to identify columns that require data wrangling. An exploratory data analysis (EDA) was performed to discover any patterns related to target distribution as well as the impact of different features on the target. After EDA, two machine learning algorithms were implemented to predict churn. Evaluation of the models was then performed based on metrics such as recall, AUC score, and confusion matrix.

EXPLORATORY DATA ANALYSIS (EDA)

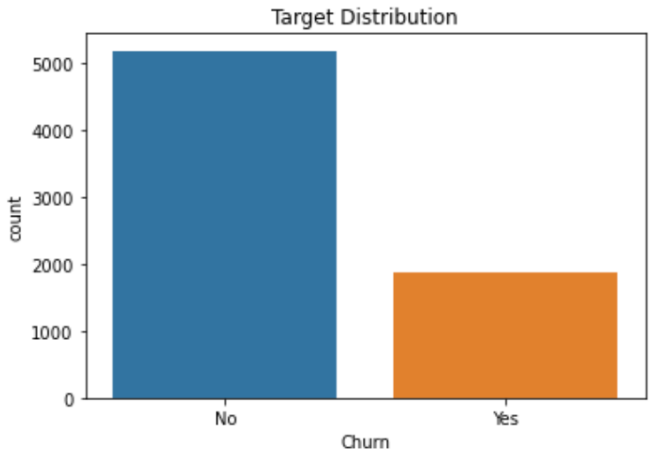
EDA was performed in four parts, with an aim to identify any patterns.

- Target distribution
- Impact of demographic features on the target
- Impact of service usage features on the target
- Impact of account features on the target

Target distribution

The target variable "Churn" contains Boolean values, where "Yes" means the customer churned and "No" means the customer did not churn. It suggests that our machine learning model

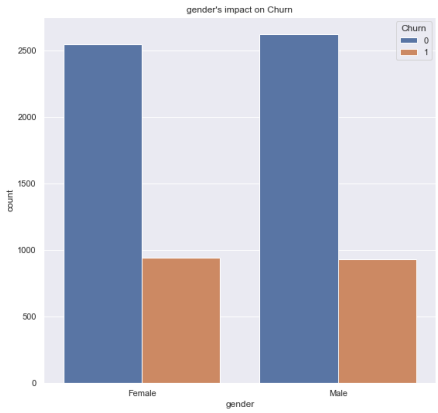
needs to solve a binary classification problem. The figure below shows an imbalanced distribution of the target label, with 5174 non-churners and 1869 churners.

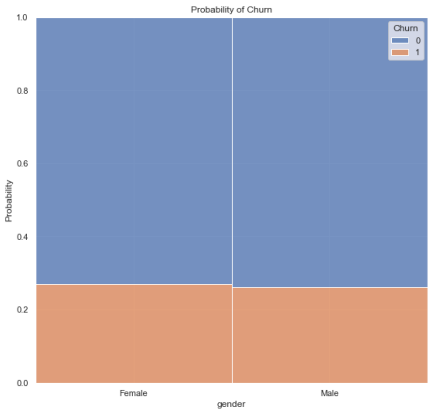


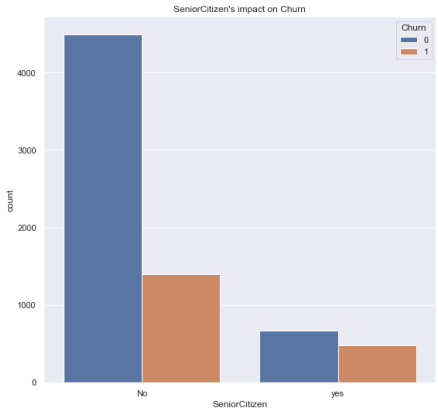
Impact of Demographic Features on Churn

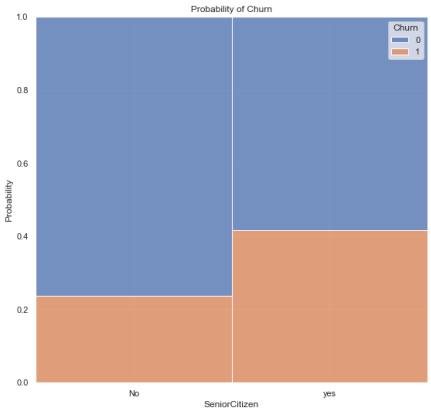
EDA was performed from two perspectives, the absolute count as shown on the left bar chart and the normalized probability as shown on the right bar chart, to explore the impact of four (4) demographic features (Gender, SeniorCitizen, Partners, Dependents) on churn. It was found that all demographic features have an obvious impact on churn probability except gender.

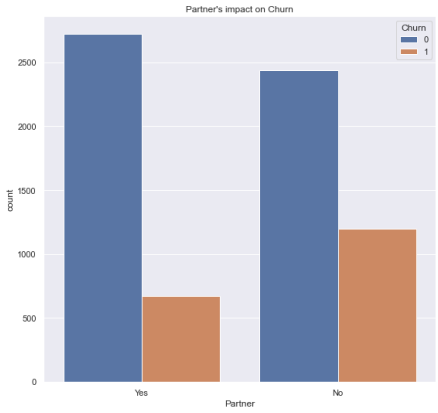
- Gender: Not much difference was found between male and female, neither in count nor in churn probability.
- SeniorCitizen: the count plot on the left showed that customers are mostly non-senior citizens; but the probability plot showed that the older people are more likely to churn.
- Partners: customers without partner(s) are more likely to churn.
- Dependents: customers without dependent(s) are more likely to churn.

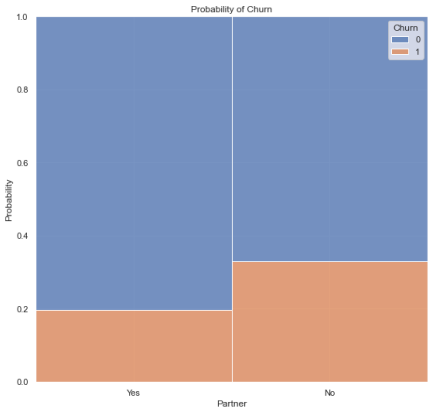

gender's impact on Churn
Churn
0
1
count
Female
Male
gender

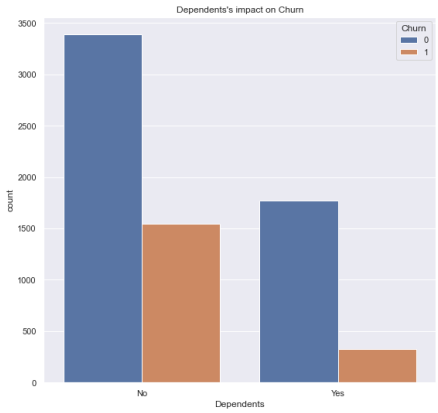

Probability of Churn
Churn
0
1
Probability
Female
Male
gender

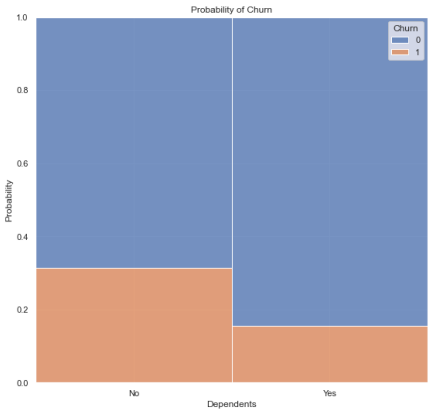

SeniorCitizen's impact on Churn
Churn
0
1
count
No
yes
SeniorCitizen


Probability of Churn
Churn
0
1
Probability
No
yes
SeniorCitizen


Partner's impact on Churn
Churn
0
1
count
Yes
No
Partner


Probability of Churn
Churn
0
1
Probability
Yes
No
Partner


Dependents's impact on Churn
Churn
0
1
count
No
Yes
Dependents

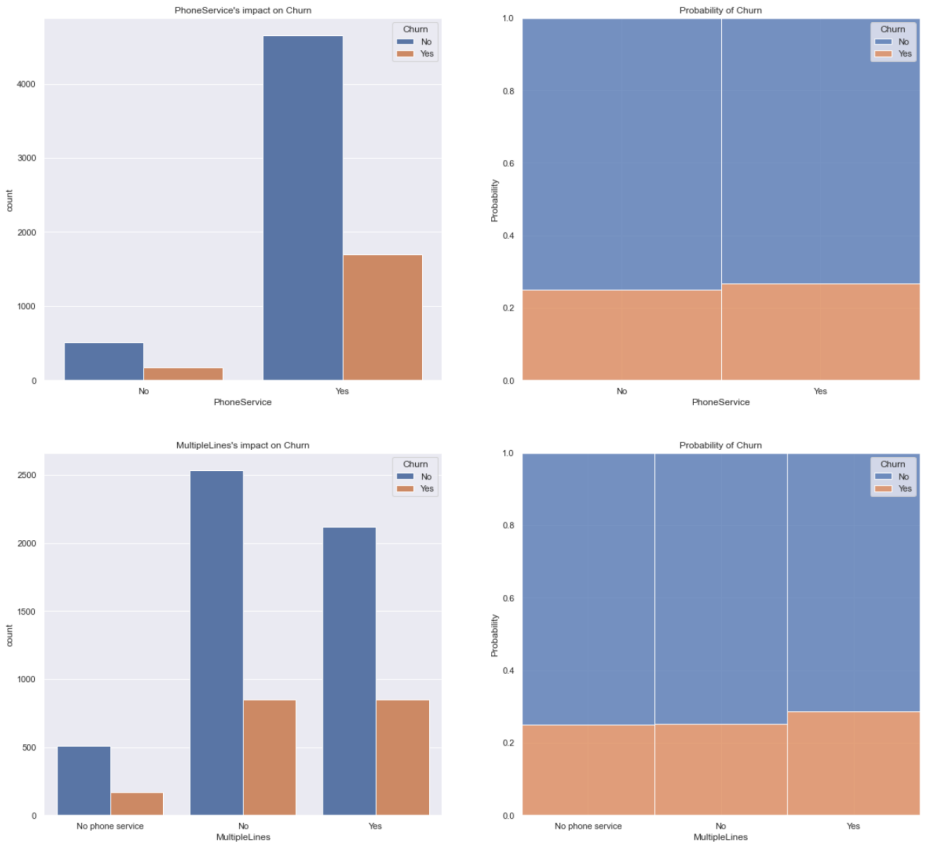

Probability of Churn
Churn
0
1
Probability
No
Yes
Dependents

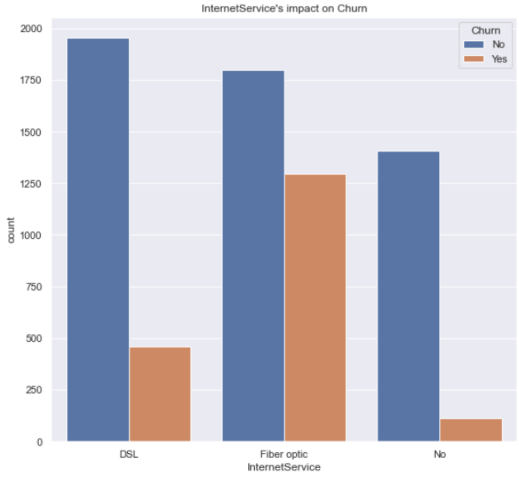
Impact of Service Usage Features on Churn

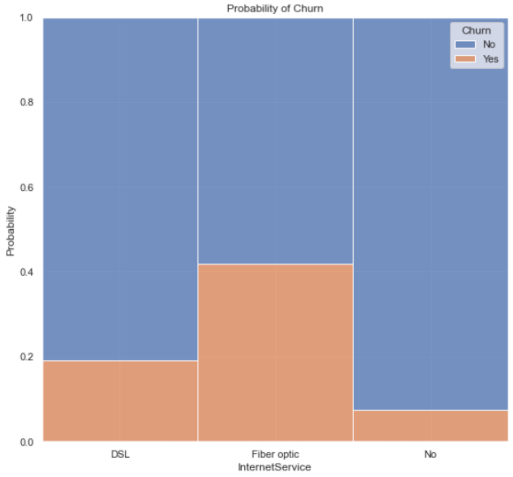
EDA was performed from two perspectives, the absolute count as shown on the left bar chart and the normalized probability as shown on the right bar chart, to explore the impact of nine (9) service usage features (‘PhoneService', 'MultipleLines', 'InternetService', 'OnlineSecurity', 'OnlineBackup', 'DeviceProtection', 'TechSupport', 'StreamingTV', 'StreamingMovies') on churn.

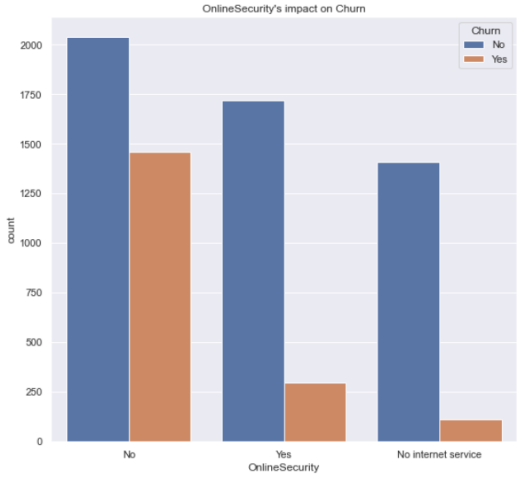
The plots below show the following insights:

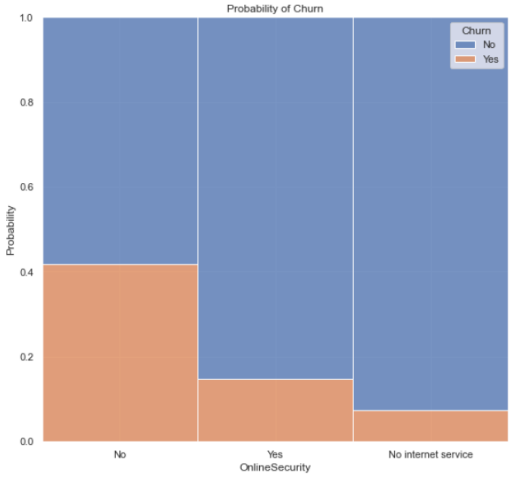
- customers who used fiber optic InternetService are twice likely to churn than those used DSL.
- customers who did not subscribe the following services are more likely to churn: 'OnlineSecurity', 'OnlineBackup', 'DeviceProtection', 'TechSupport'
- Subscription of the following services has little impact on churn probability: 'PhoneService', 'MultipleLines', 'StreamingTV','StreamingMovies'

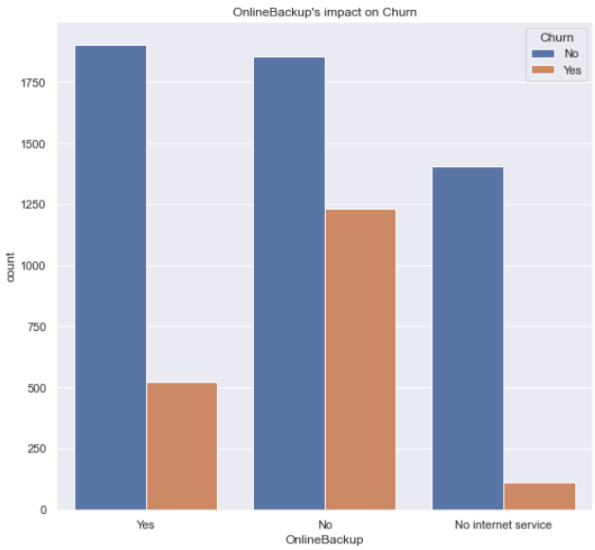


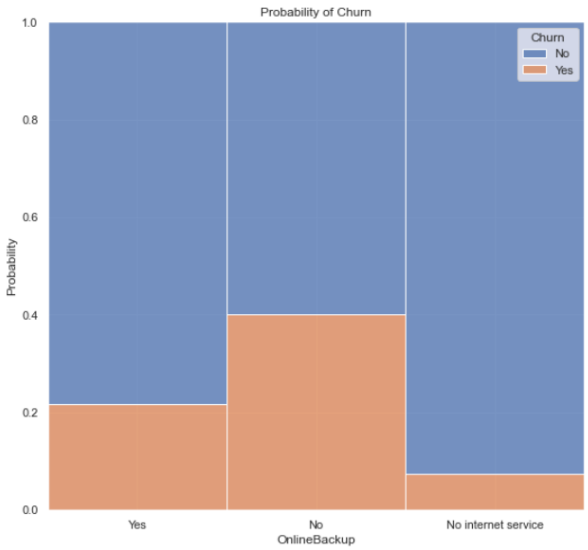

InternetService's impact on Churn
Churn
No
Yes
count
2000
1750
1500
1250
1000
750
500
250
0
DSL
Fiber optic
No
InternetService

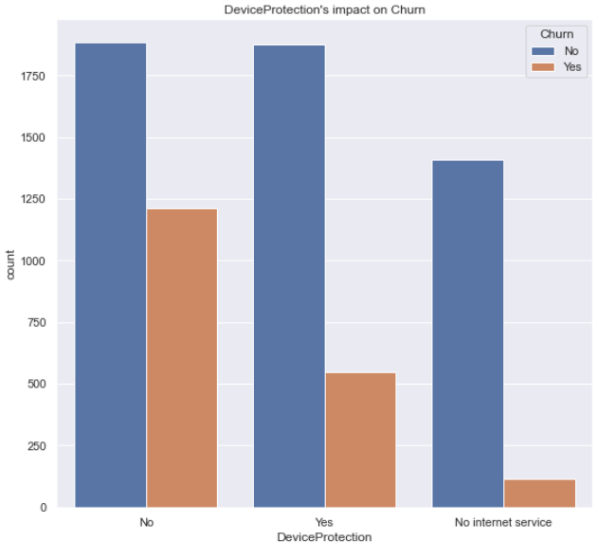

Probability of Churn
Churn
No
Yes
Probability
1.0
0.8
0.6
0.4
0.2
0.0
DSL
Fiber optic
No
InternetService

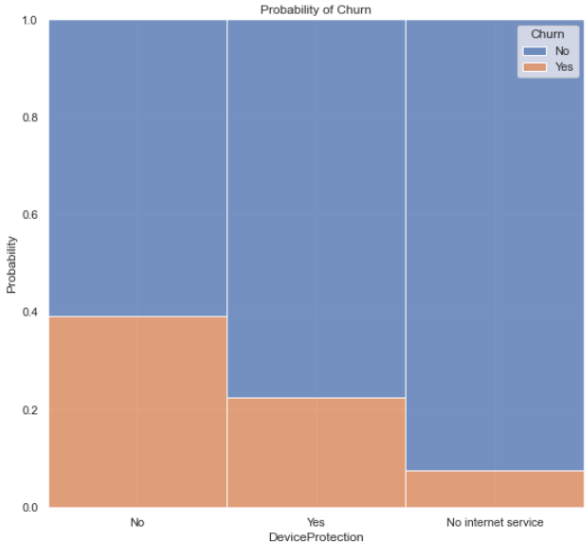

OnlineSecurity's impact on Churn
Churn
No
Yes
count
2000
1750
1500
1250
1000
750
500
250
0
No
Yes
No internet service
OnlineSecurity

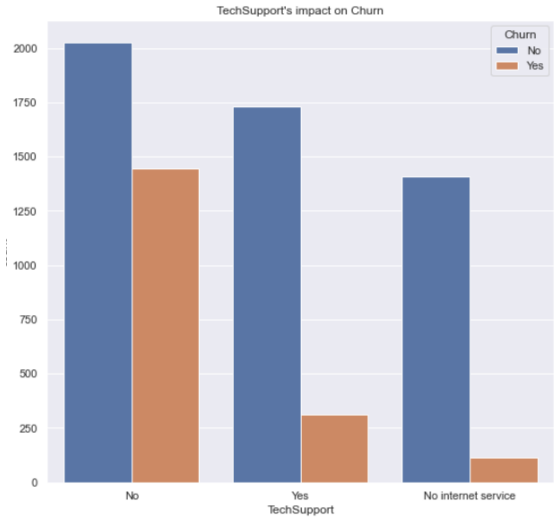

Probability of Churn
Churn
No
Yes
Probability
1.0
0.8
0.6
0.4
0.2
0.0
No
Yes
No internet service
OnlineSecurity

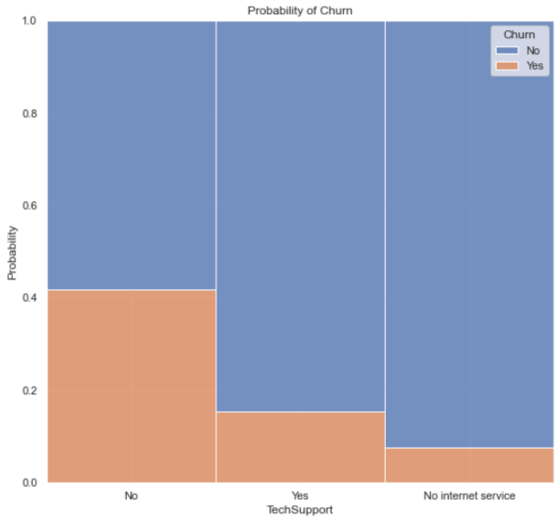

OnlineBackup's impact on Churn
Churn
No
Yes
count
0
250
500
750
1000
1250
1500
1750
Yes
No
No internet service
OnlineBackup

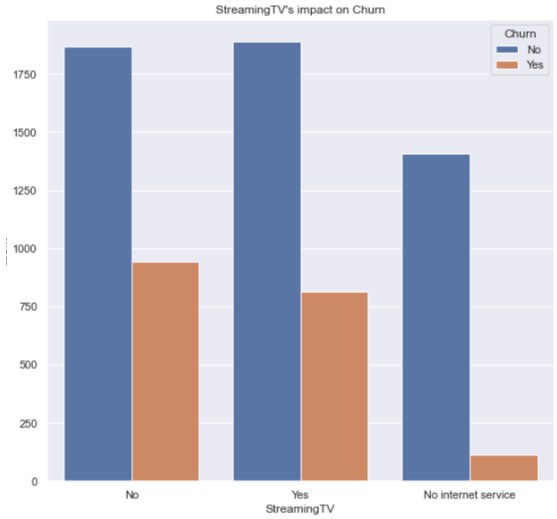

Probability of Churn
Churn
No
Yes
Probability
0.0
0.2
0.4
0.6
0.8
1.0
Yes
No
No internet service
OnlineBackup

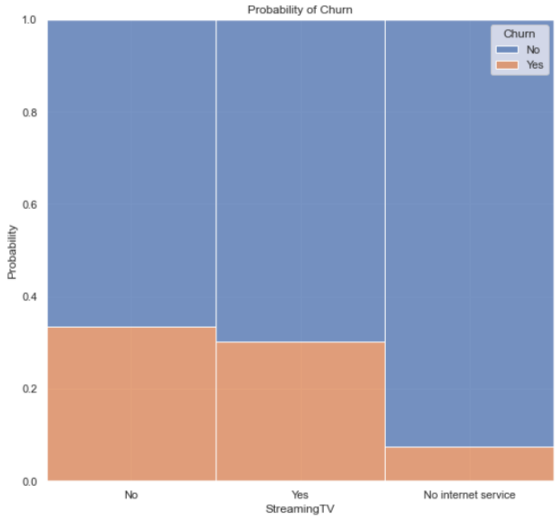

DeviceProtection's impact on Churn
Churn
No
Yes
count
0
250
500
750
1000
1250
1500
1750
No
Yes
No internet service
DeviceProtection

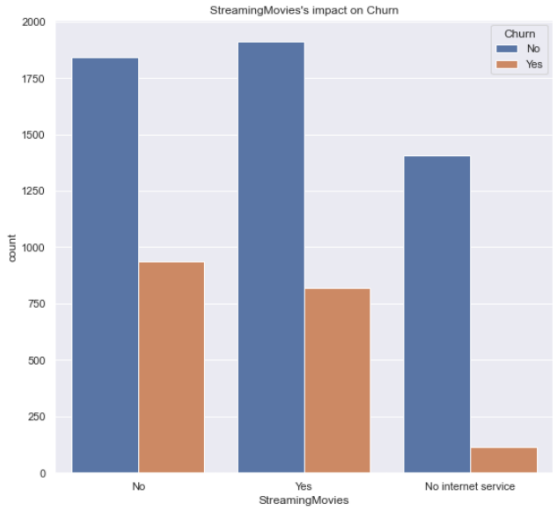

Probability of Churn
Churn
No
Yes
Probability
0.0
0.2
0.4
0.6
0.8
1.0
No
Yes
No internet service
DeviceProtection

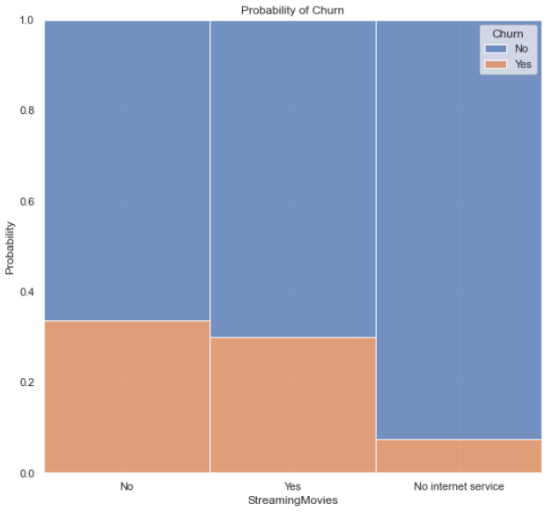

TechSupport's impact on Churn
Churn
No
Yes
2000
1750
1500
1250
1000
750
500
250
0
No
Yes
No internet service
TechSupport


Probability of Churn
Churn
No
Yes
1.0
0.8
0.6
0.4
0.2
0.0
Probability
No
Yes
No internet service
TechSupport


StreamingTV's impact on Churn
Churn
No
Yes
1750
1500
1250
1000
750
500
250
0
No
Yes
No internet service
StreamingTV


Probability of Churn
Churn
No
Yes
1.0
0.8
0.6
0.4
0.2
0.0
Probability
No
Yes
No internet service
StreamingTV


StreamingMovies's impact on Churn
Churn
No
Yes
2000
1750
1500
1250
1000
750
500
250
0
count
No
Yes
No internet service
StreamingMovies

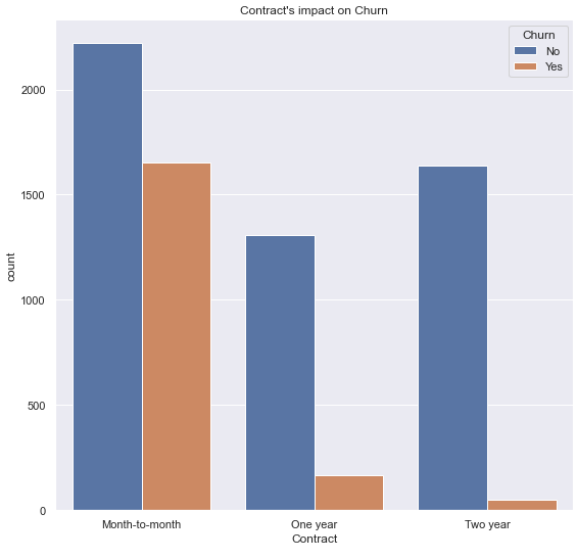

Probability of Churn
Churn
No
Yes
1.0
0.8
0.6
0.4
0.2
0.0
Probability
No
Yes
No internet service
StreamingMovies

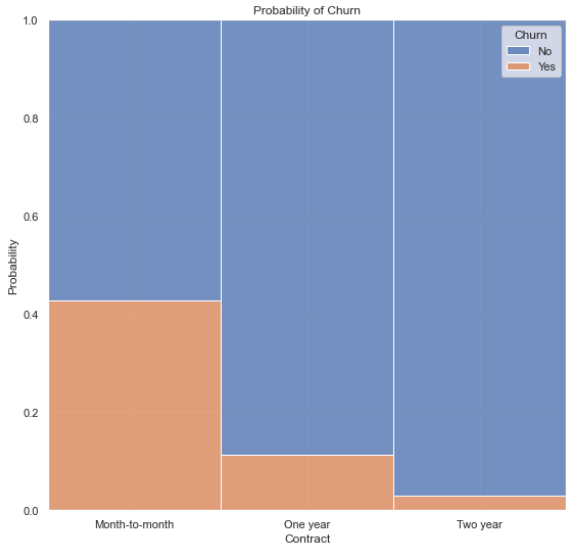
Impact of Account Features on Churn

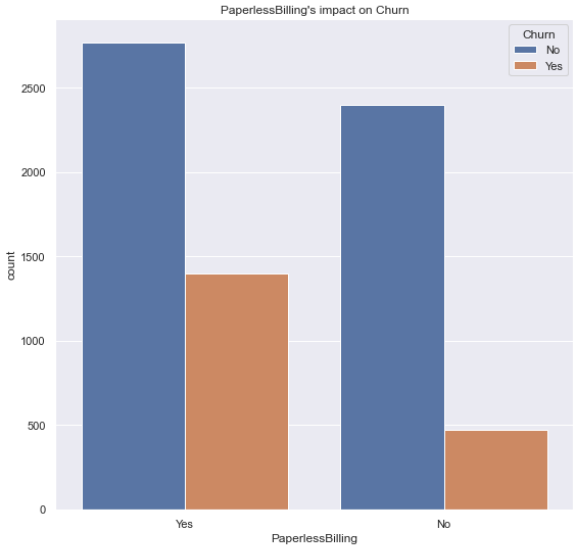
Account data can be divided into three(3) categorical features ("Contract", "PaperlessBilling", "PaymentMethod") and three(3) numeric features ("tenure", "MonthlyCharges", "TotalCharges"). For categorical features, the EDA was performed from two perspectives, the absolute count as shown on the left bar chart and the normalized probability as shown on the right bar chart. For numeric features, the EDA was performed from the perspective of churn probability only.

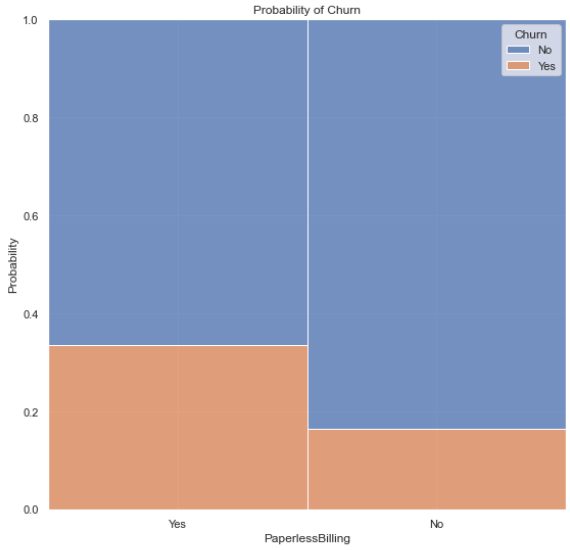
The plots of the categorical features below produced the following insights:

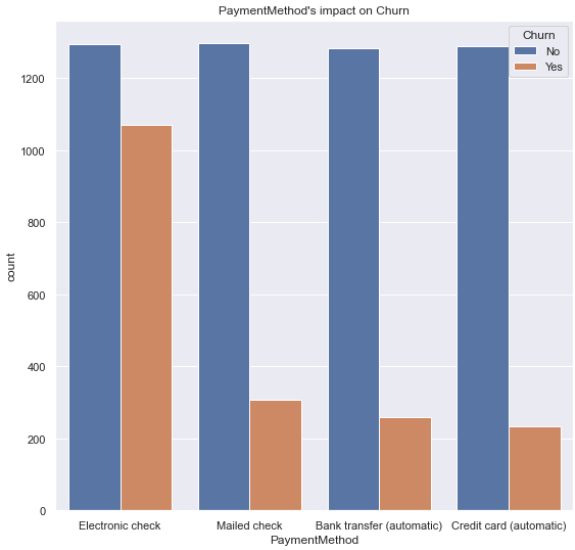
- "Contract": the count plot on the left showed that the company has a lot of month-to-month customers, and the probability plot on the right showed that those monthly-based customers are far more likely to churn than customers in 1-year or 2-year contract.
- "PaperlessBilling": customers who signed up for paperless billing has a higher probability to churn.
- "PaymentMethod": customers who make payments by electronic check are about twice likely to churn than people who use other payments such as mailed check, bank transfer and credit card.

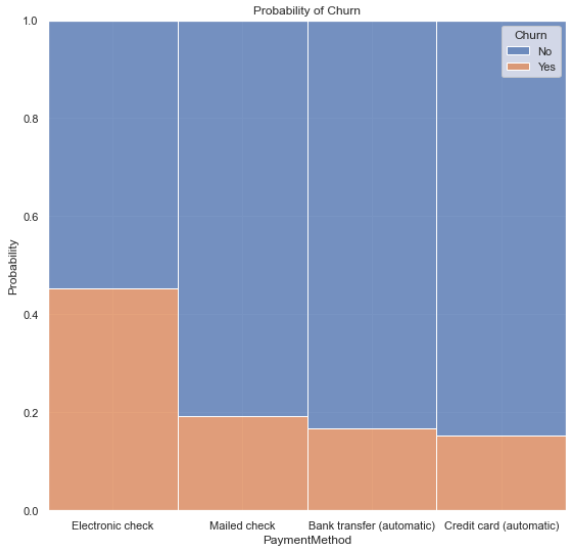

Contract's impact on Churn
Churn
No
Yes
count
2000
1500
1000
500
0
Month-to-month
One year
Two year
Contract


Probability of Churn
Churn
No
Yes
Probability
1.0
0.8
0.6
0.4
0.2
0.0
Month-to-month
One year
Two year
Contract


PaperlessBilling's impact on Churn
Churn
No
Yes
count
2500
2000
1500
1000
500
0
Yes
No
PaperlessBilling

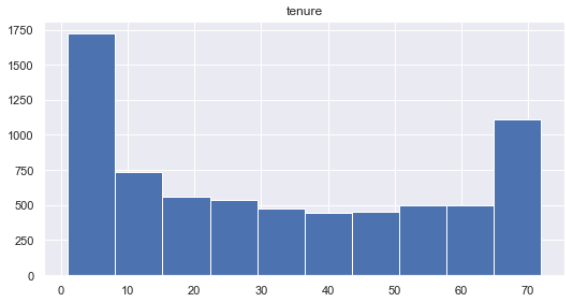

Probability of Churn
Churn
No
Yes
Probability
1.0
0.8
0.6
0.4
0.2
0.0
Yes
No
PaperlessBilling

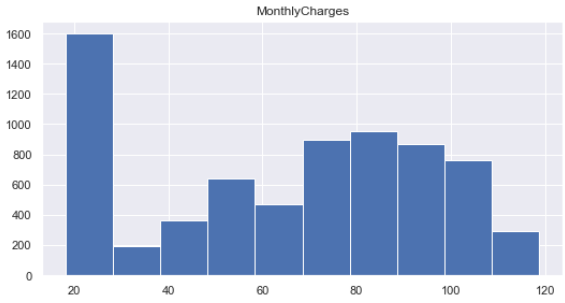

PaymentMethod's impact on Churn
Churn
No
Yes
count
1200
1000
800
600
400
200
0
Electronic check
Mailed check
Bank transfer (automatic)
Credit card (automatic)
PaymentMethod

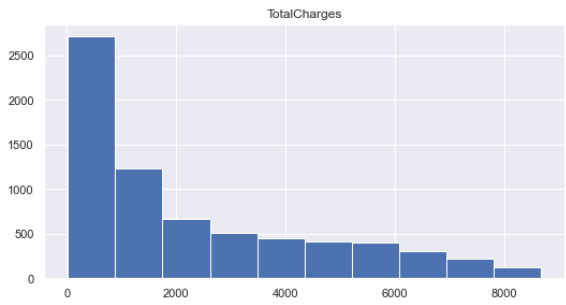

Probability of Churn
Churn
No
Yes
Probability
1.0
0.8
0.6
0.4
0.2
0.0
Electronic check
Mailed check
Bank transfer (automatic)
Credit card (automatic)
PaymentMethod

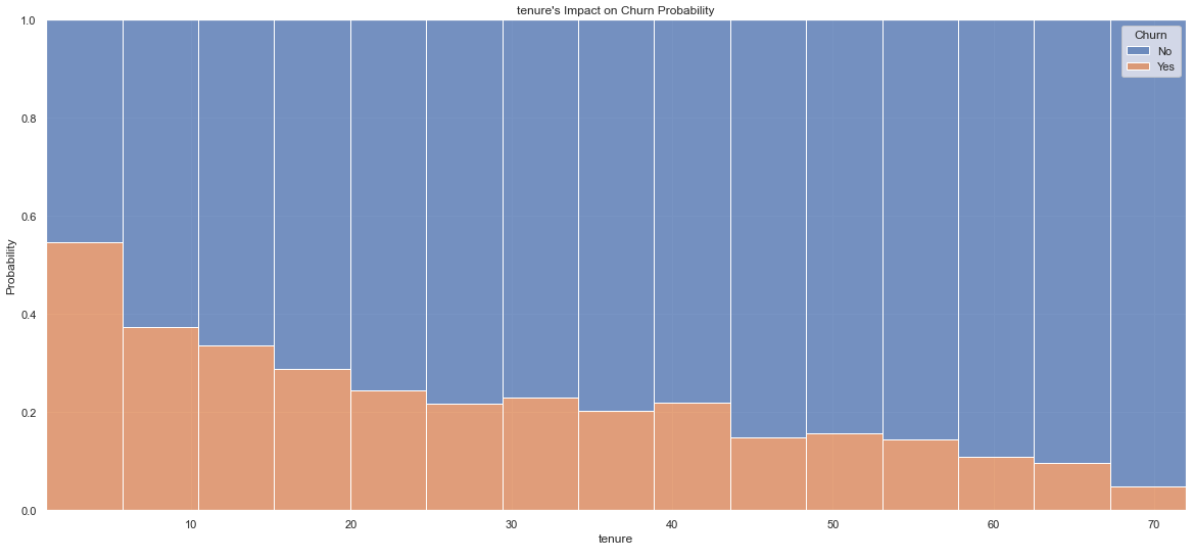
The plots of numeric features below yielded the following insights:

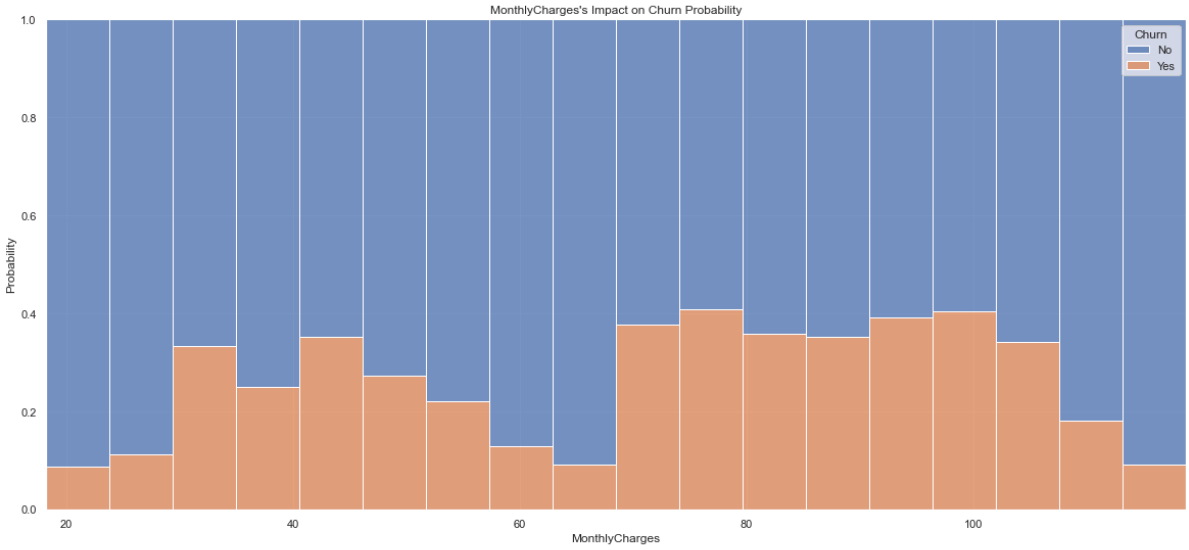
- "Tenure": tenure seems to be negatively correlated with churn, which is logical as customers who have stayed with the company for a longer time are usually less likely to churn.
- "MonthlyCharges": no clear pattern except that customers who paid very little or very much are less likely to churn.
- "TotalCharges": one unexpected insight was revealed: total charges is negatively correlated with churn. It may be because loyal customers subscribe a lot more services and thus incur higher charges.

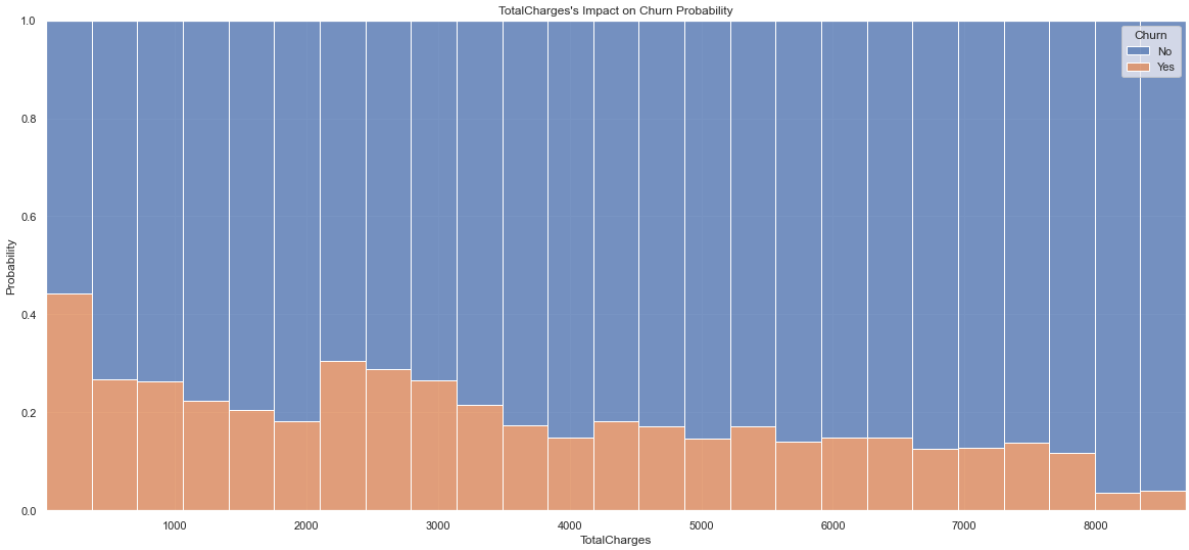







tenure's Impact on Churn Probability
Churn
No
Yes
Probability
1.0
0.8
0.6
0.4
0.2
0.0
10
20
30
40
50
60
70
tenure


MonthlyCharges's Impact on Churn Probability
Churn
No
Yes
Probability
1.0
0.8
0.6
0.4
0.2
0.0
20
40
60
80
100
MonthlyCharges


TotalCharges's Impact on Churn Probability
Churn
No
Yes
Probability
1.0
0.8
0.6
0.4
0.2
0.0
1000
2000
3000
4000
5000
6000
7000
8000
TotalCharges

PREDICTIVE MODELING

Data Preparation

Data was prepared for modeling, such as encoding all categorical features using OneHotEncoder and scaling all numerical features using StandardScaler. I used the ColumnTransformer and the Pipeline tools from sklearn to make the process more streamlined.

Model Training

The prepared dataset was then split into training set and test set. The training dataset was used to train two machine learning algorithms, which were random forest and logistic regression. I also used class_weight="balanced" since the EDA above showed the target label was imbalanced.

Model Evaluation

The two models were evaluated using metrics such as recall, f1-score, ROC AUC score, and confusion matrix.

Classification report (to check recall and f1-score)

classification report (model 1 - Random Forest):

```
y_pred_rf = rf.predict(X_test)
```

```
print(classification_report(y_test, y_pred_rf))
```

```
              precision    recall  f1-score   support

           0       0.82      0.90      0.86      1549
           1       0.63      0.45      0.53       561

    accuracy                           0.78      2110
   macro avg       0.73      0.68      0.69      2110
weighted avg       0.77      0.78      0.77      2110
```

classification report (model 2 - Logistic Regression):

```
y_pred_LR = LR.predict(X_test)
```

```
print(classification_report(y_test, y_pred_LR))
```

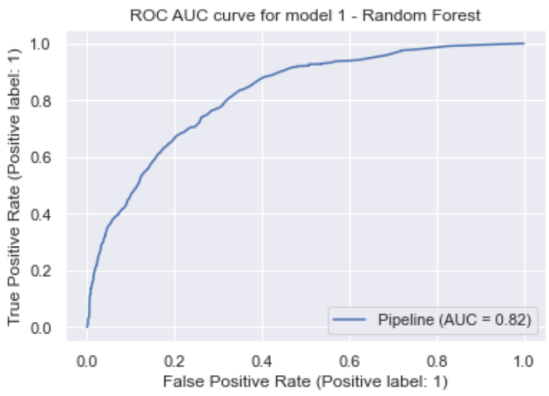
```
              precision    recall  f1-score   support

           0       0.91      0.72      0.80      1549
           1       0.51      0.81      0.63       561

    accuracy                           0.74      2110
   macro avg       0.71      0.76      0.72      2110
weighted avg       0.81      0.74      0.76      2110
```

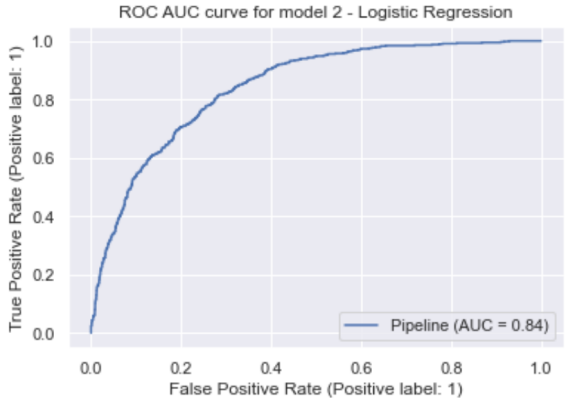
Since my main interest is to identify the churners (label 1) instead of the non-churners (label 0), I would focus on prediction of label "1" (churners); precision and recall are always contradictory but as misclassifying a churner into non-churners is more costly than the other way around, I would attach more importance to recall than to precision. The classification report showed that random forest performs not as well as logistic regression in terms of "recall" and "fl-score" for label 1(churn): random forest has a recall

of 0.45, much lower than 0.81 from the logistic regression. Logistic regression's f1-score 0.63 also outperforms that of random forest (0.53). Therefore, in terms of both recall and f1-score, logistic regression is the winner.

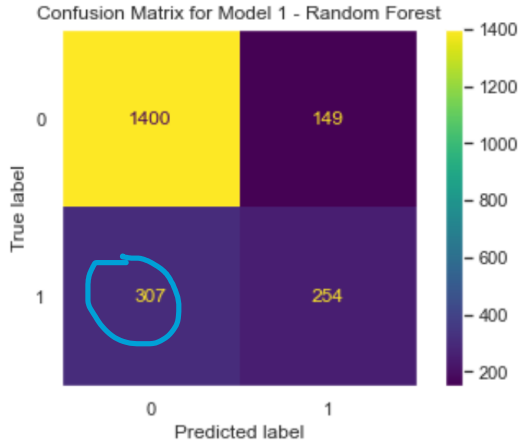
ROC AUC

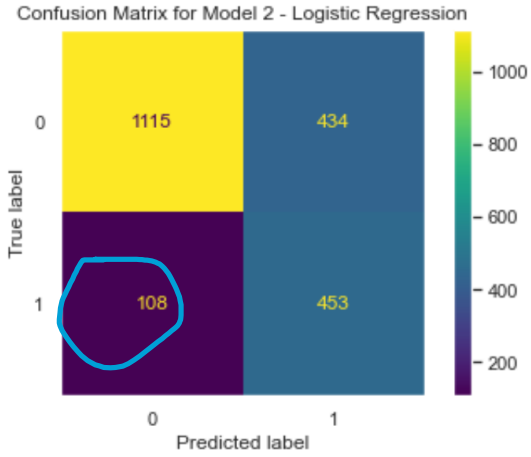
Logistic Regression's ROC AUC score is 0.84, slightly higher than that of random forest (0.82).





Confusion matrix





The confusion matrix revealed that model 2 (logistic regression) is able to identify churners much better than random forest, with only 108 cases missed, about one third (1/3) of the misclassification cases by random forest (307).

Based on the above evaluation, it's concluded that model 2 (logistic regression) outperforms model 1 (random forest) for this dataset and should be selected for churn prediction. This AI algorithm helps to boost the churn prediction rate from its base rate of 0.27% (1869 churners / 7043 total customers = 0.27) to 0.81% (recall), offering great value to the CRM department for designing client retention programs.

CONCLUSION

This project compared two machine learning algorithms - random forest classifier and logistic regression classifier – in predicting whether a customer will churn or not. The logistic regression model is the winner for this dataset based on performance metrics such as recall, f1-score, ROC AUC score, and confusion matrix. The project has successfully demonstrated that machine learning algorithms can be applied to reduce risk for an organization. Prediction of customer churn in telecom industry helps the firm identify potential churners in advance and then target these customers with effective retention programs, thus reducing the churn rate and minimizing the risk of profit loss.

BIBLIOGRAPHY

Burez, J., Van den Poel, D., Ngai, E., & Ying, W. (2009). Customer churn prediction using improved balanced random forests. *Expert Systems With Applications*, *36*(3), 5445–5449. https://doi.org/10.1016/j.eswa.2008.06.121

Dalvi, P. K., Khandge, S. K., Deomore, A., Bankar, A., & Kanade, V. A. (2016). Analysis of customer churn prediction in telecom industry using decision trees and logistic regression. *2016 Symposium on Colossal Data Analysis and Networking (CDAN)*. https://doi.org/10.1109/cdan.2016.7570883

Senthan, P., Rathnayaka, R., Kuhaneswaran, B., & Kumara, B. (2021). Development of Churn Prediction Model using XGBoost - Telecommunication Industry in Sri Lanka. *2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*. https://doi.org/10.1109/iemtronics52119.2021.9422657

Ullah, I., Raza, B., Malik, A. K., Imran, M., Islam, S. U., & Kim, S. W. (2019). A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector. *IEEE Access*, *7*, 60134–60149. https://doi.org/10.1109/access.2019.2914999

Umayaparvathi, V., & Iyakutti, K. (2016). A Survey on Customer Churn Prediction in Telecom Industry: Datasets, Methods and Metrics. *International Research Journal of Engineering and Technology (IRJET)*, *Volume: 03*(Issue: 04). https://d1wqtxts1xzle7.cloudfront.net/54561005/IRJET-V3I4213-with-cover-page-v2.pdf?Expires=1669665317&Signature=Du7WcI8DCFW~qsk63gcEExJCZGQ0mEhQAZKyTTcTnTmkXDBwcBpqy-mU~G~~WNmPHvxxOrZgxDIzuedtIQh3g9PJbB6io2ojfj7TxXwGouOZxl8D8Cpvxh0XJoRp~wX1xshB-Q6Z5EWk9-S50PPzej45NqIvyc7NQ7UnrzaCuzb54bxvNE1qiPX9vhNsqfz9gco7Xl2TvTxx-PnRnFFZGI82OYliaYnZHRoD9mnj~tYb1ZWeMQBFSv~o5nnYqENftIwbh9vj5bQcf1ijJM4viaAgeWPYsc9JvFBlgbXU7YkfahE6h93L8PFrCisMyzcp60TfaRWMPSFofoDLldtMSw__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA

Zhang, T., Moro, S., & Ramos, R. F. (2022). A Data-Driven Approach to Improve Customer Churn Prediction Based on Telecom Customer Segmentation. *Future Internet*, *14*(3), 94. https://doi.org/10.3390/fi14030094