Capstone 2 Project Report:
Telco Customer Churn

Problem statement:

The objective of this project is to predict customer behavior in order to retain them. The goal is to analyze all relevant customer data and develop an algorithm that can model churn (customer retention index) with at least 80% accuracy. Also, by understanding the reasons behind customers leaving, focused customer retention programs can be developed to increase retention chances.

Context:

Telco is a fictional phone and internet services company providing services to its customers in California. Telco has collected data from a total of 7043 customers which includes multiple important customer demographics, services that each customer pays for, a satisfaction score and a ‘churn’ score. The churn score shows how long the customer has been using the services of Telco. The executives at Telco want to understand how accurately can the Churn feature can be predicted based on other features. The executives would like to see at least 80% accuracy in prediction. This will help understand which customers are more likely to leave and what specific actions can be taken to increase the chances of retention.

Data:

There is one source dataset which contains information from 7043 customers (rows) and 21 features (columns). First the data will be analyzed to understand how many features are usable for modelling. In order to understand the relationship between different features and ‘churn’ correlation map will be generated. The dataset has to be divided into a training set (‘train’) and testing set (‘test’). Different models will be derived using ‘train’ and the performance will be tested on the ‘test’.

Data Wrangling:

The original dataset was loaded and it constituted of data from 7043 different customers and 21 columns describing the customer data such as demographics, services they opted for, their billing details etc. and there was one column called ‘Churn’ which had a binary value depending on whether the customer churned or not. Each of these customer descriptive columns will be called a feature and in the data wrangling stage each feature will be analyzed to see their usability and/or if any anomalies or outliers are present.

Looking at the datatype for each feature revealed that there are 3 numerical features namely: ‘SeniorCitizen’, ‘Tenure’ and ‘MonthlyCharges’ and rest were categorical with datatype as ‘object’. Upon investigation it was found that ‘SeniorCitizen’ only had binary values (1 or 0), representing ‘Yes’ or ‘No’. Among the categorical columns ‘TotalCharges’ had numerical values stores as text. This feature was converted to a numerical datatype (float) and treated as a numerical feature and “SeniorCitizen’ as a categorical feature.

Next null value were checked and there were 11 null values for the ‘TotalCharges’ feature. No other feature had any null values. So, this dataset was mostly clean. Next it was investigated if we could infill the missing values for ‘TotalCharges’ by multiplying the ‘Tenure’ with ‘MonthlyCharges’. Unfortunately, the ‘Tenure’ feature for those 11 instances had 0 value. So, infill for those 11 null values was not possible. Eventually, the 11 instances (rows) were dropped from the dataset.

There was one feature called customer ID which was unique for each customer. This feature was completely dropped as this will not help predict the churn behavior in any way. After data cleaning we had 7032 customers and 20 features out of which 3 were numerical and 17 were categorical. Next, box plots were generated for the numerical shows which are shown in figure 1 to check for outliers and there were no outliers for any of those three features.

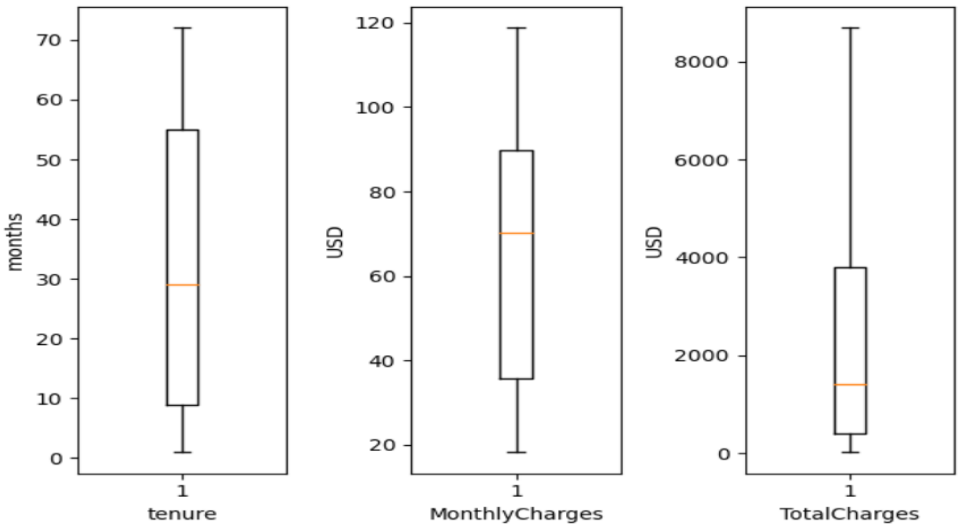


Figure 1: Boxplots for the numerical features in the data to check for outliers.

Exploratory Data Analysis:

After cleaning the data, the relationship between individual features and ‘Churn’ was explored in detail. There are a total of 20 features after data cleaning of which one is ‘Churn’ which is the variable we are trying to predict. EDA explored the relationship between other 19 features and ‘Churn’. First the overall distribution of customers who Churned vs non churned was explored. Below is a figure that shows the overall distribution of customers with respect to ‘Churn’.

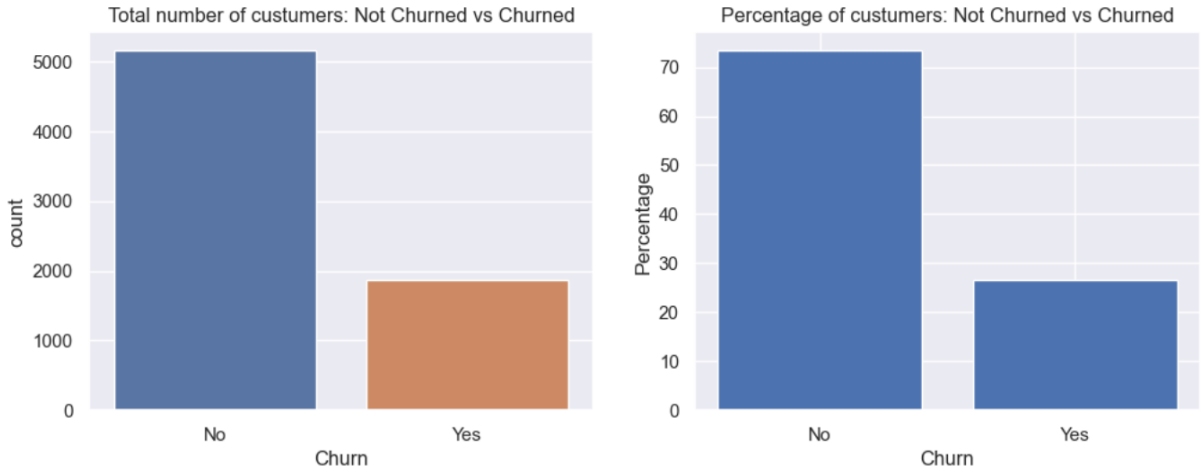
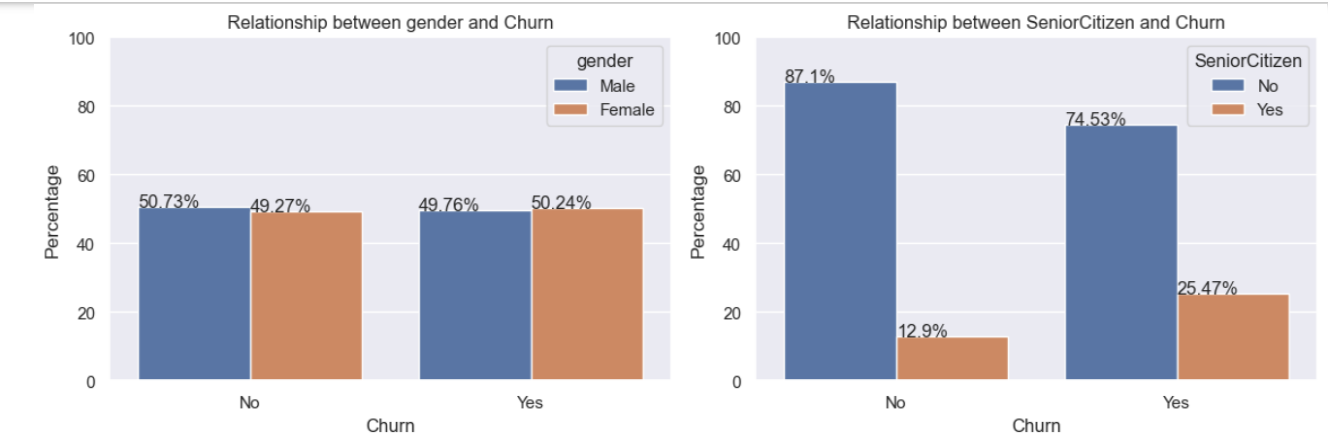


Figure 2: Churn numbers and percentages of all customers

For easier understanding, the features are grouped based on their meaning. The various groups of features are as follows:

- Customer demographics
  - Gender
  - Senior Citizen
  - Partner
  - Dependents
- Customer services
  - Phone Service
  - Multiple Lines
  - Internet Service
  - Online Security
  - Online Backup
  - Device Protection
  - Tech Support
  - Streaming TV
  - Streaming Movies
- Customer payment information
  - Contract
  - Paperless Billing
  - Payment Method
- Customer billing
  - Tenure
  - Monthly Charges
  - Total Charges



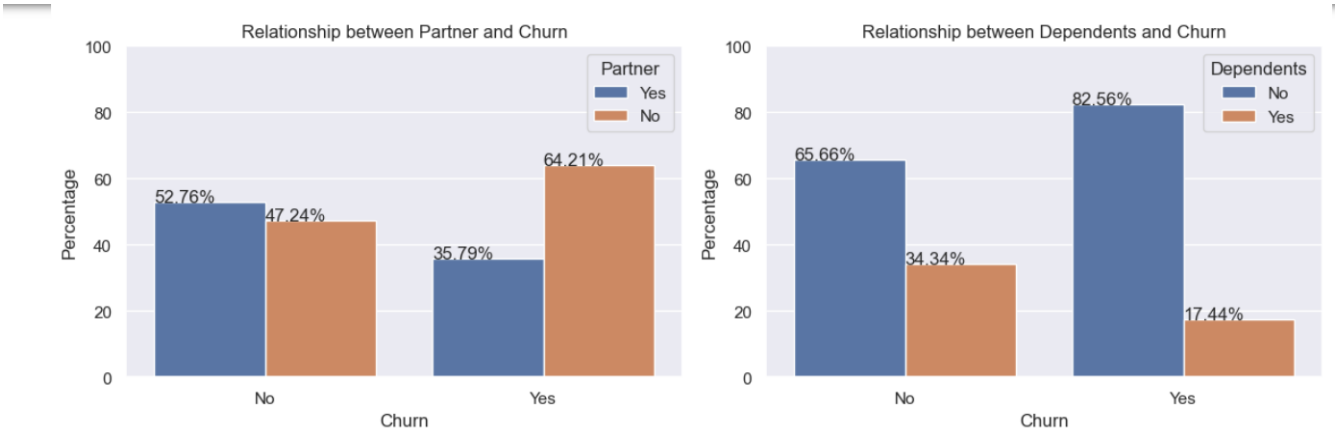
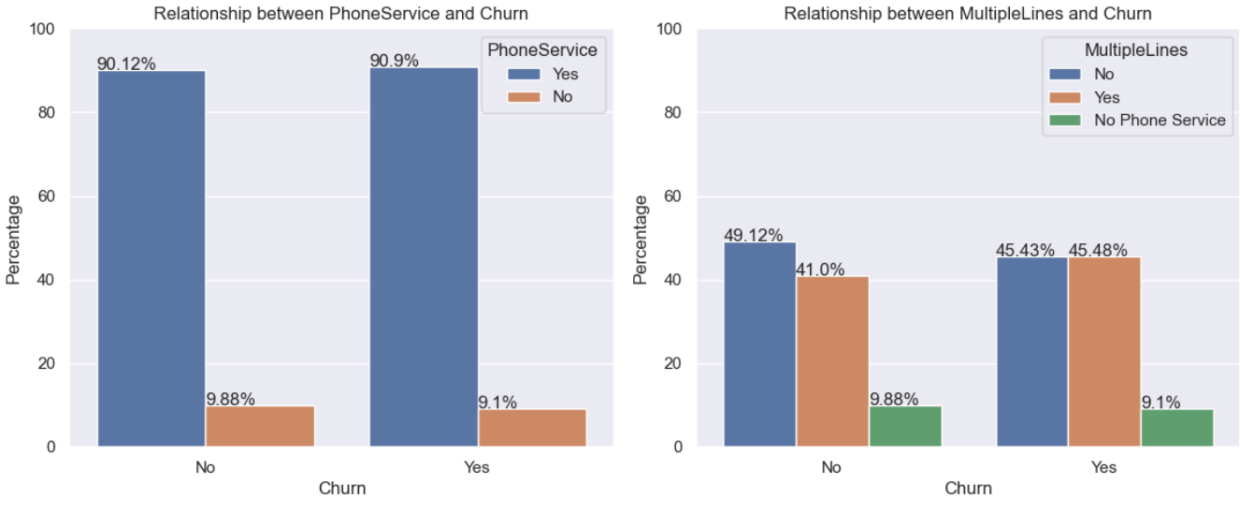
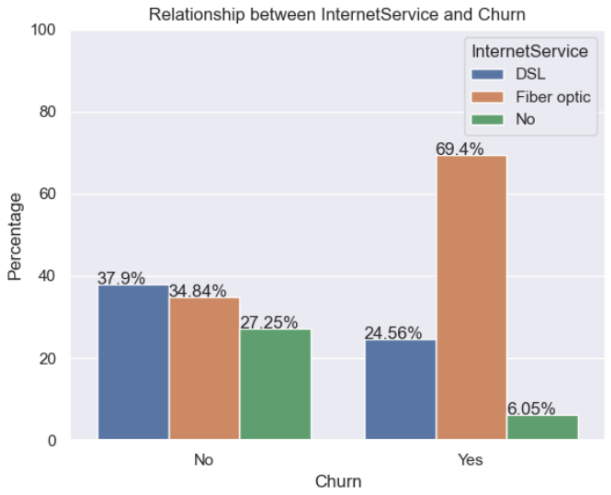


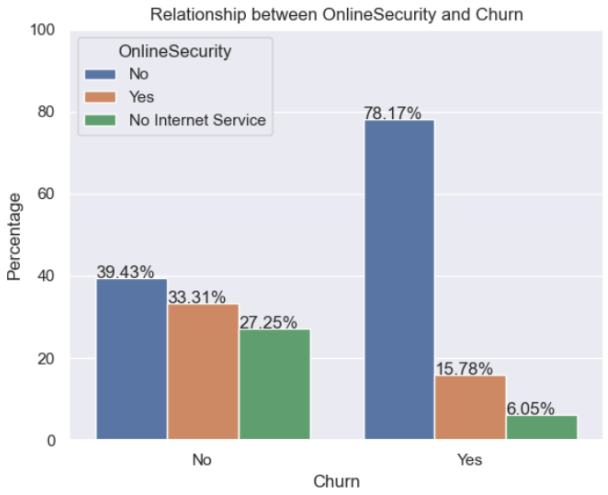
Figure 3: EDA for features representing customer demographics

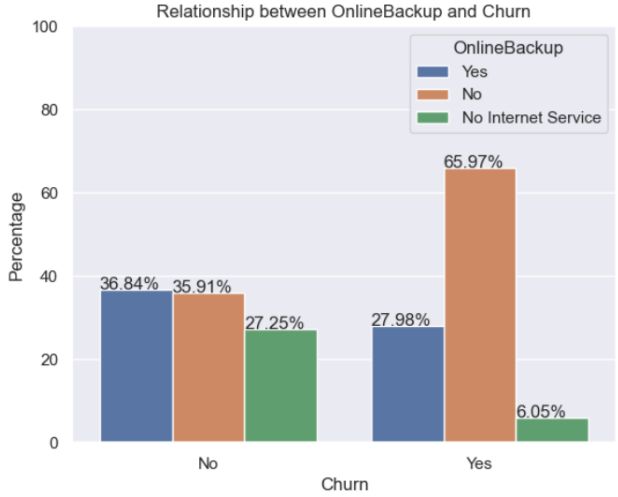
EDA for Customer demographics:

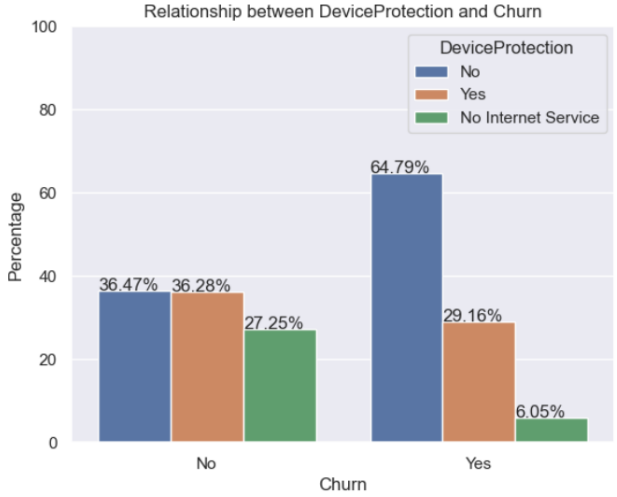
- **Gender:** There is no clear pattern between customer's gender and churn.
- **Senior Citizen:** It seems like being a senior citizen increases the chances of churning.
- **Partner:** Customers without a partner are more likely to churn.
- **Dependents:** Most customers do not have dependents. Customers without dependents are more likely to churn.

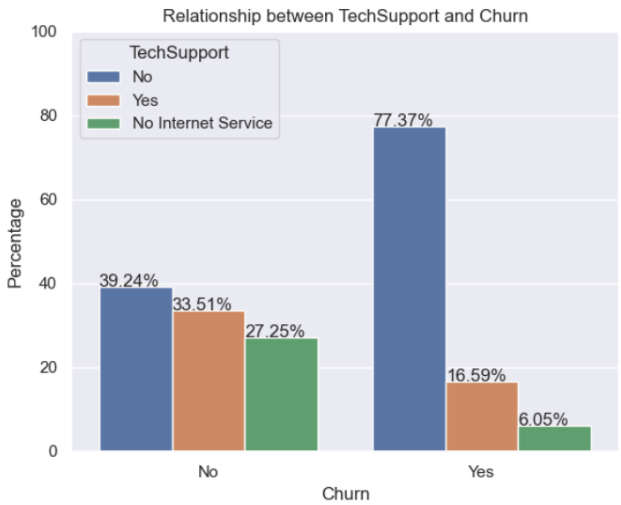


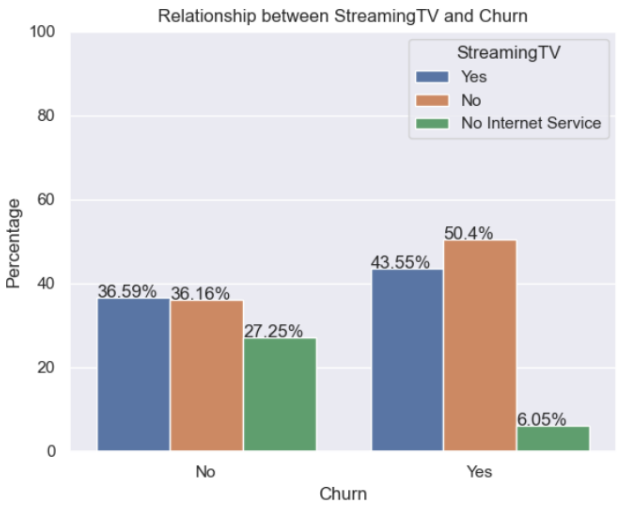

Relationship between InternetService and Churn
InternetService
DSL
Fiber optic
No
Percentage
Churn
No
Yes
37.9%
34.84%
27.25%
24.56%
69.4%
6.05%


Relationship between OnlineSecurity and Churn
OnlineSecurity
No
Yes
No Internet Service
Percentage
Churn
No
Yes
39.43%
33.31%
27.25%
78.17%
15.78%
6.05%


Relationship between OnlineBackup and Churn
OnlineBackup
Yes
No
No Internet Service
Percentage
Churn
No
Yes
36.84%
35.91%
27.25%
27.98%
65.97%
6.05%


Relationship between DeviceProtection and Churn
DeviceProtection
No
Yes
No Internet Service
Percentage
Churn
No
Yes
36.47%
36.28%
27.25%
64.79%
29.16%
6.05%


Relationship between TechSupport and Churn
TechSupport
No
Yes
No Internet Service
Percentage
Churn
No
Yes
39.24%
33.51%
27.25%
77.37%
16.59%
6.05%


Relationship between StreamingTV and Churn
StreamingTV
Yes
No
No Internet Service
Percentage
Churn
No
Yes
36.59%
36.16%
27.25%
43.55%
50.4%
6.05%

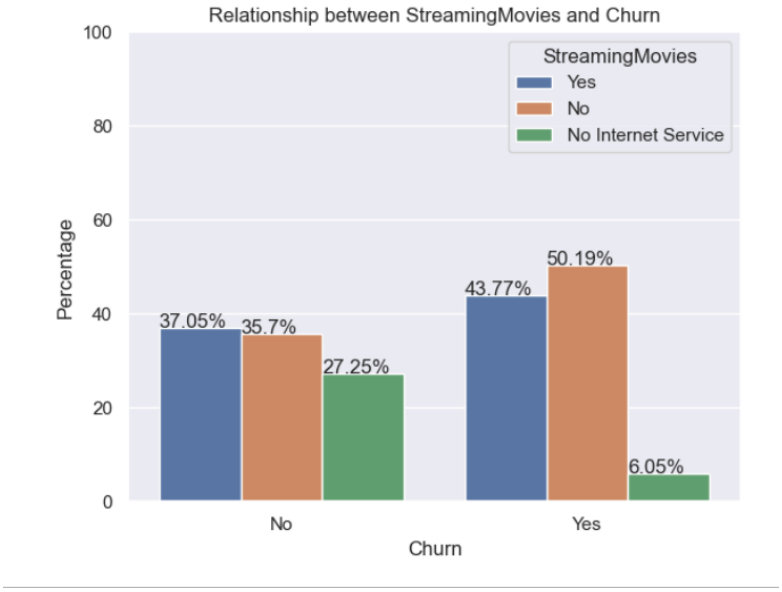


Figure 4: EDA for features representing customer services

EDA for Customer Services:

- **Phone service:** Most of the customers (roughly 90%) chose phone service and this feature seems to have no apparent relationship with churn. This feature may not have much impact in predicting whether a customer would churn and can be dropped to make a prediction model simpler but this will be evaluated in the modelling stage.
- **Multiple lines:** Among the customers who chose phone service, customers who did not choose multiple lines are less likely to churn by a small percentage.
- **Internet service:** Customers who chose the fiber optic are highly likely to churn. This is an indication that the fiber optic service provided by the company needs to be evaluated and compared with what the competitors have to offer.
- **Online security:** Customers that did not choose online security service option are more likely to churn. Again, something to investigate for the company. Did customers without online security have poor performance issues or spam threats etc.
- **Online backup:** This feature shows a similar distribution as for 'Online security'. Again, customers with no online backup option are more likely to churn. It may be interesting to see if these two features are related. Correlation heatmap generated later can be used to verify if the above-mentioned assumption is true.
- **Device protection and Tech support:** Both these features show a similar relationship with churn as 'Online security' and 'Online backup'. Customer who chose no as more likely to Churn. Later we will see if there is any correlation among these 4 features.
- **Streaming TV and Streaming movies:** Customers who chose no for these services are slightly more likely to churn but the relationship is not as strong as with the previous 4 services. These two streaming related services show a similar distribution and their correlation will the checked later.

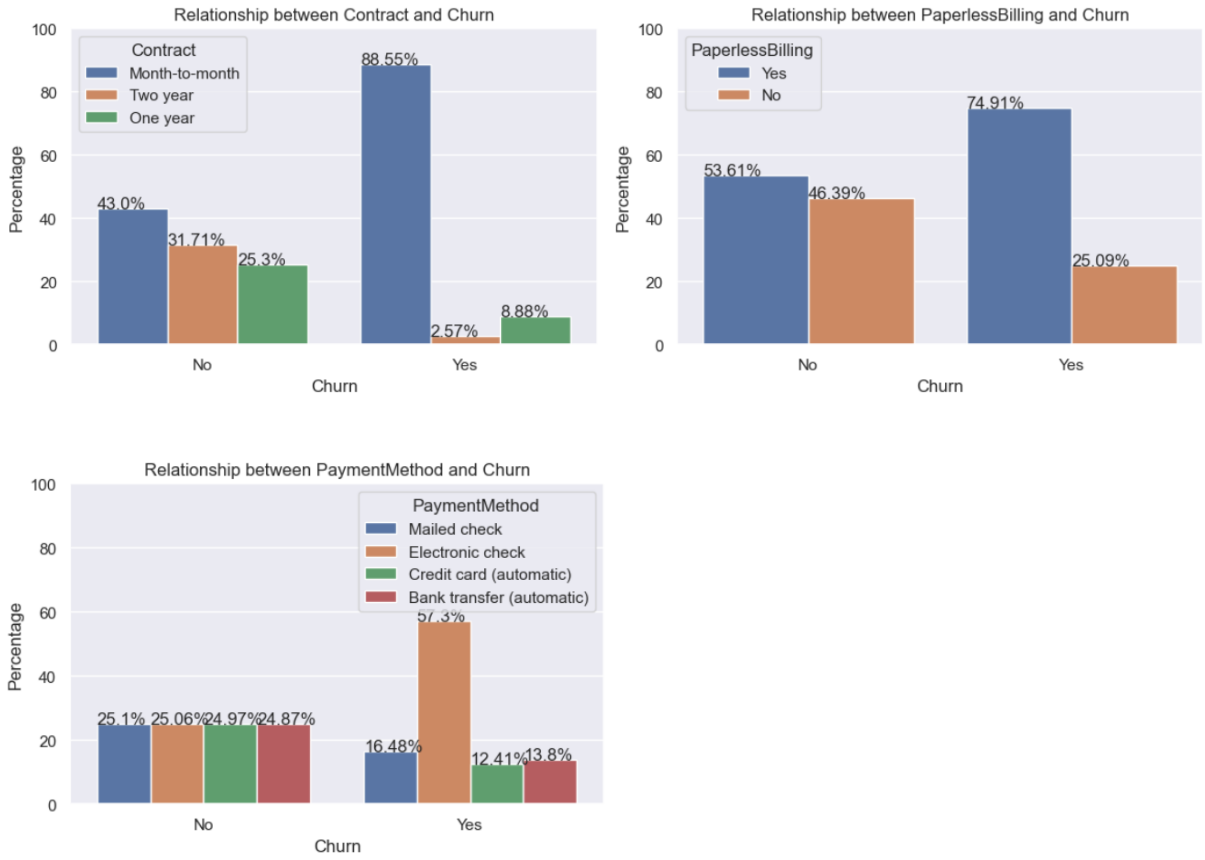


Figure 5: EDA for features representing customer payment information

EDA for Customer payment information:

- **Contract:** Customers who were on month-to-month were more likely to churn than the ones in longer 1 year or 2 year contracts. This is probably as expected.
- **Paperless billing:** Customers who chose paperless billing option were more likely to churn. This does not make sense and maybe we can check if this feature has a strong correlation with any other feature that can explain this relationship with churn.
- **Payment method:** Customers who chose electronic check option are also highly likely to churn compared to other payment types. Maybe e-check service has some processing issues/delays. Needs to be investigated.

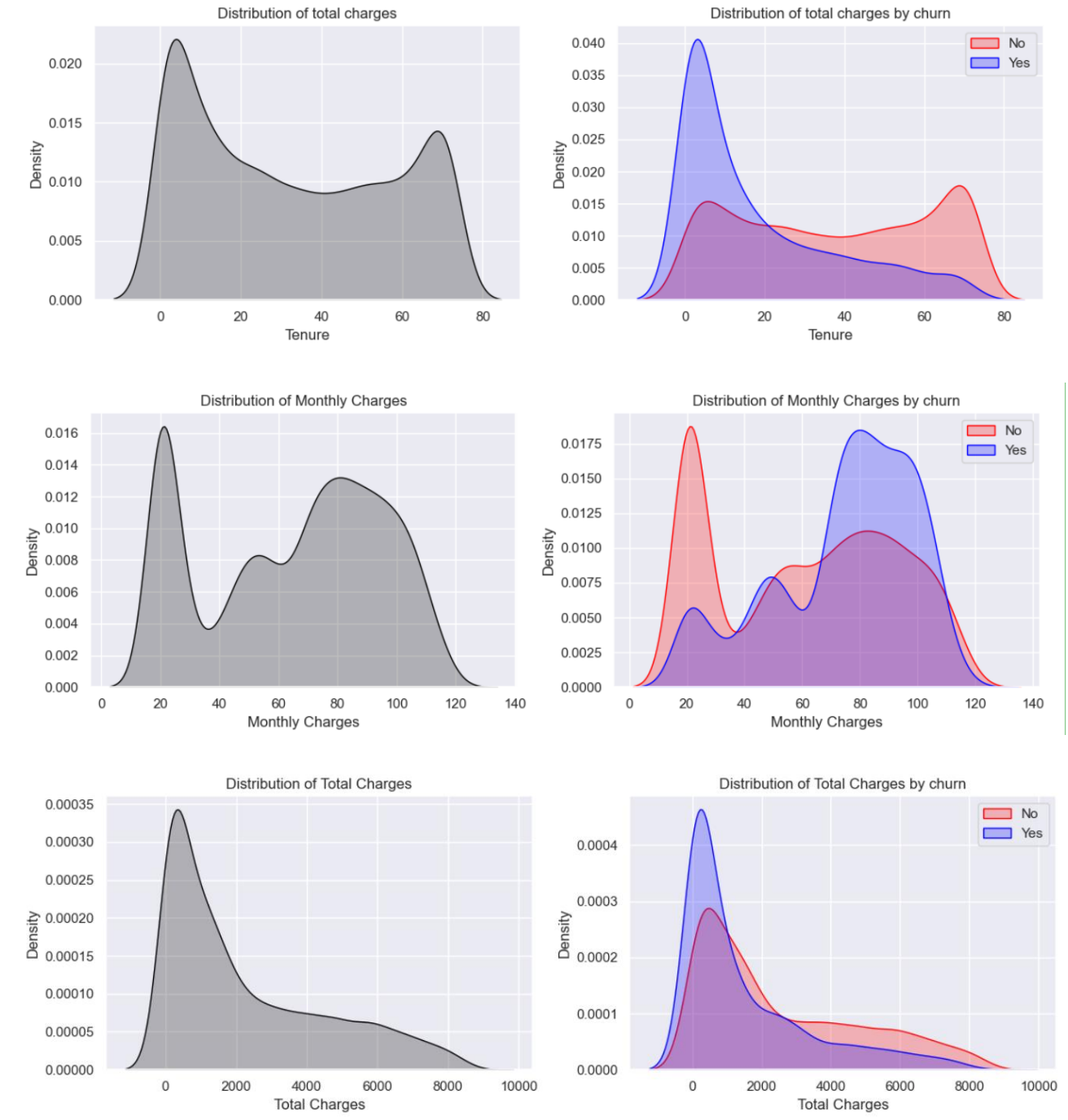


Figure 6: EDA for features representing customer billing information

EDA for Billing information:

- **Tenure:** For all the data this feature shows a bimodal distribution with 2 peaks (at around 3 and 70). Seprating the customers by churn and looking at the distributions explains the 2 peaks. Customers who churned stayed for a short period and customers who did not churn stayed longer with the company.

- **Monthly charges:** For all data the distribution was again bimodal with two peaks at around 20 and 80. By plotting for customers based on churn we can see that customers who churned on an average pai much higher than the customers who did not churn. High costs of service seems to be a strong reason behind customers to churn.
- **Total charges:** We can that the distribution is not normal and more like chi-square with a long tail. It is not clear if the total charges has a strong relationship with churn.

Correlation between different features:
Looking at the correlation heatmap in figure 7, let us use a threshold of >0.6 and <-0.6 to select features that have strong correlation.

- **Tenure and contract:** 0.68
- **Tenure and total charges:** 0.83 Tenure and contract are correlated which is expected. Customers choosing month-to-month have likely shorter tenure compared to customers who are on yearly contract. Total charges likely is a function on monthly charges and tenure, so it is expected to have a strong correlation coefficient.
- **Phone service and Multiple Lines:** 0.67 Phone service is likely is redundant feature. Customers who did not choose phone service couldn't have chosen any other features such as multiple lines, online backup etc.
- **Internet service, Online security, Online backup, Device protection, Tech support, Streaming TV and Streaming movies:** correlation higher than 0.58 From the plots shown earlier we suspected there could be strong correlation between the service features and the correlation heatmap confirms the suspicion. It seems like the same customers opted for similar kind of services.
- **Monthly charges vs Different services:** correlation > 0.64 Monthly charges depends on number of services chosen and hence we see a strong correlation for the services and monthly charges.
- **Partner and dependents:** correlation = 0.45 Customers with partners more likely to have dependents (children). Interesting correlation between the two features.

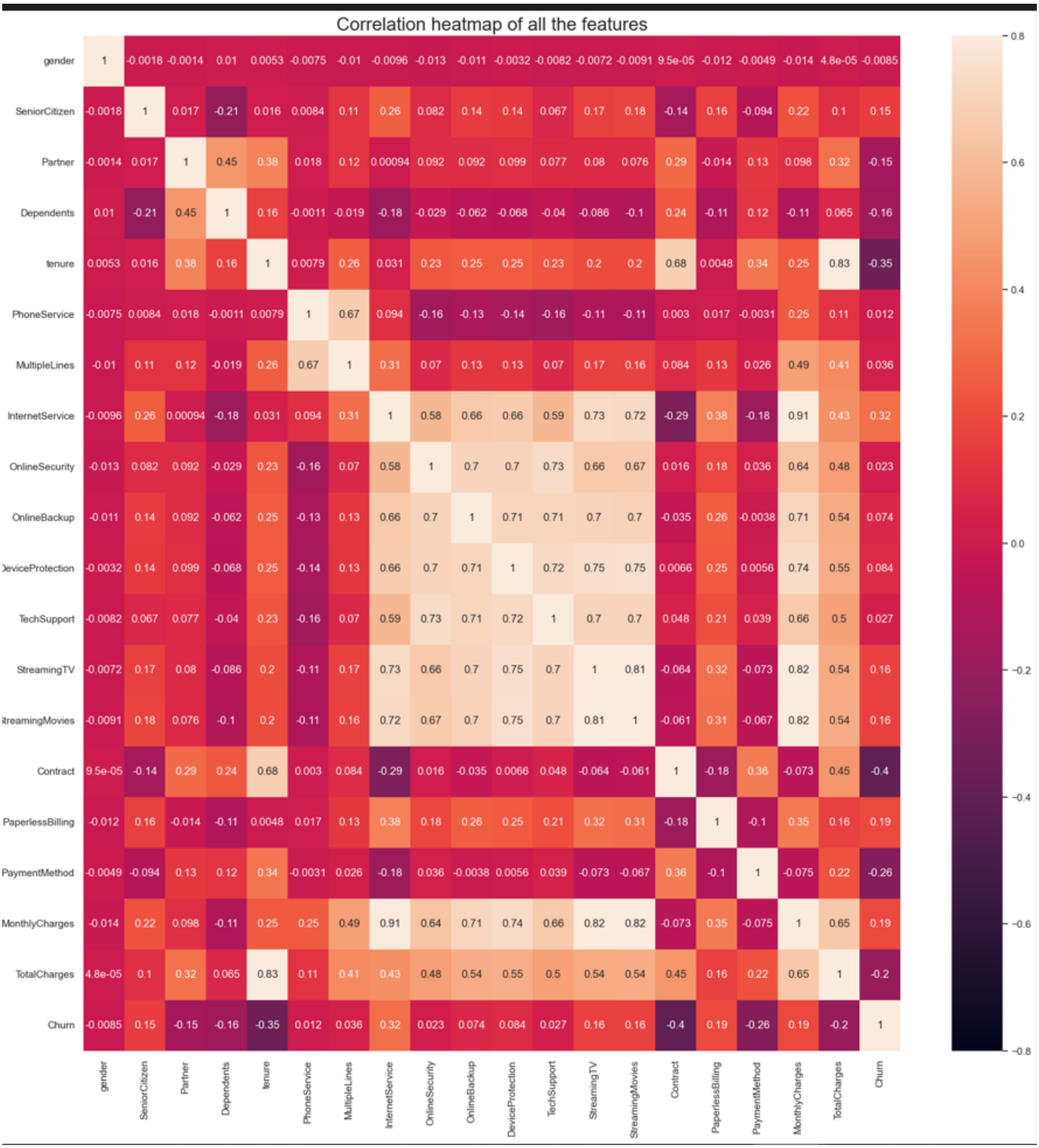


Figure 7: Correlation map between different features

Data Preprocessing:

One-hot encoding of categorical features: The categorical features were one-hot encoded using the ‘get_dummies’ function. The 16 categorical features resulted in 28 one-hot encoded features. The ‘Churn’ column was dropped.

Split into training and testing test: The data were then split into training set and testing set with 30% of the data reserved for testing purposes. While splitting the data ‘stratify’ option was used to keep the inherent percentages of ‘Churn’ and ‘Not churn’ consistent across the training and testing data. The percentages were confirmed to be consistent.

Scaling of numerical features: The numerical features were scaled such that their mean was 0 and standard deviation 1. The scalers were derived from the training data only and applied to both training and test data to preserve the independence of test data.

Infilling for imbalanced groups: Since our data has imbalanced groups as shown in figure 2, another set of training data was generated by infilling the less represented feature (customers that churned). SMOTE was used to infill the data. After applying it was verified that both classes has equal representation.

Modeling Results:

Modeling 1: Using imbalanced training data and default modeling parameters. The different models that will be trained and tested are:

- Logistic Regression without any regularization
- Decision Tree
- Random Forest
- XGBoost
- LightGBM
- Gradient Boost
- ADA Boost
- Support vector machine
- K-nearest neighbors

The metrics for evaluation are: Accuracy, Precision, Recall, F1 SCore, ROC-AUC

	Model	Accuracy	Precision	Recall	F1 Score	AUC-ROC
0	Logistic Regression	0.809479	0.670968	0.556150	0.608187	0.728688
1	Decision Tree	0.735071	0.501754	0.509804	0.505747	0.663230
2	Random Forest	0.789573	0.642336	0.470588	0.543210	0.687844
3	XGBoost	0.780569	0.603376	0.509804	0.552657	0.694218
4	LightGBM	0.796209	0.645880	0.516934	0.574257	0.707144
5	Gradient Boosting	0.799052	0.655329	0.515152	0.576846	0.708512
6	ADA Boosting	0.801896	0.652452	0.545455	0.594175	0.720113
7	Support Vector Machine	0.801896	0.688654	0.465241	0.555319	0.694531
8	K-Nearest Neighbors	0.754976	0.543651	0.488414	0.514554	0.669965

Figure 8: Results from Modeling 1

Figure 8 summarizes the results from modeling 1. Generally, the precision score is higher than recall. This may be because of imbalanced classes. F1-score is the most reliable measure for evaluating the models. Based on the F1 scores the best models are Logistic regression and ADABoost. KNN and single decision tree gave the worst results. These results are as expected, however it is interesting to see that support vector machine's performance is lower that Logistic regression and boosting algorithms. Next, we will run the same models using the infilled data using SMOTE.

Modeling 2: Using balanced training data and default modeling parameters. The different models that will be trained and tested are:

- Logistic Regression without any regularization
- Decision Tree
- Random Forest
- XGBoost
- LightGBM
- Gradient Boost
- ADA Boost
- Support vector machine
- K-nearest neighbors

The metrics for evaluation are: Accuracy, Precision, Recall, F1 SCore, ROC-AUC

	Model	Accuracy	Precision	Recall	F1 Score	AUC-ROC
0	Logistic Regression	0.758294	0.532484	0.745098	0.621100	0.754085
1	Decision Tree	0.718483	0.475842	0.579323	0.522508	0.674103
2	Random Forest	0.774408	0.568438	0.629234	0.597293	0.728109
3	XGBoost	0.754028	0.530086	0.659537	0.587768	0.723894
4	LightGBM	0.768246	0.550992	0.693405	0.614049	0.744378
5	Gradient Boosting	0.756872	0.530000	0.755793	0.623071	0.756528
6	ADA Boosting	0.754028	0.524533	0.800357	0.633733	0.768803
7	Support Vector Machine	0.746445	0.515815	0.755793	0.613160	0.749427
8	K-Nearest Neighbors	0.707109	0.467429	0.729055	0.569638	0.714108

Figure 9: Results from Modeling 2

After SMOTE it seems like the recall scores are higher than precision. Overall accuracy is lower but the F1 score is higher. The best classifiers are ADAboost, Gradient boosting, Logistic Regression and Support vector machine in that order. Next, we will try different parameters for ADABoost, Logistic Regression, Random Forest and SVM.

Summary of hyperparameter tuning:

Random Forest Classifier best paramters: {'max_depth': 20, 'n_estimators': 200}
Random Forest Classifier default paramters: {'max_depth': none, 'n_estimators': 100}

ADA Boost Classifier best parameters: {'learning_rate': 0.5, 'n_estimators': 300} ADA Boost Classifier default parameters: {'learning_rate': 1, 'n_estimators': 50}

Support Vector Machine Classifier best parameters: {'C': 10, 'gamma': 1} Support Vector Machine Classifier deafult parameters: {'C': 1, 'gamma': 'scale'}

Logistic Regression Classifier best paramters: {'C': 1, 'penalty': 'l2'} Logistic Regression Classifier default paramters: {'C': 1, 'penalty': 'l2'}

We will build the RondomForest, ADABoost and SVM using the best parameters and evaluate using the test data. The default parameters of the Logistic Regression classifier are the best from our hyperparameter tuning test.

	Model	Accuracy	Precision	Recall	F1 Score	AUC-ROC
0	Random Forest Tuned	0.774408	0.566929	0.641711	0.602007	0.732089
1	ADA Boosting Tuned	0.751659	0.521893	0.786096	0.627312	0.762641
2	Support Vector Machine Tuned	0.749289	0.530303	0.499109	0.514233	0.669503

Figure 10: Results from tuned models

It seems like there is only a marginal difference in the performance after using the tuned parameters. Suggests that for this dataset the default parameters are quite robust. Based on our results from default settings ADABoost gave us the best F1 score: 0.634. We will use this classifier to plot the confusion matrix and the classification report.

Final Model selected: ADABoost with default parameters.

Final model accuracy = 0.754
Final model precision = 0.525
Final model recall = 0.800
Final model f1-score = 0.634
Final model auc-roc = 0.769

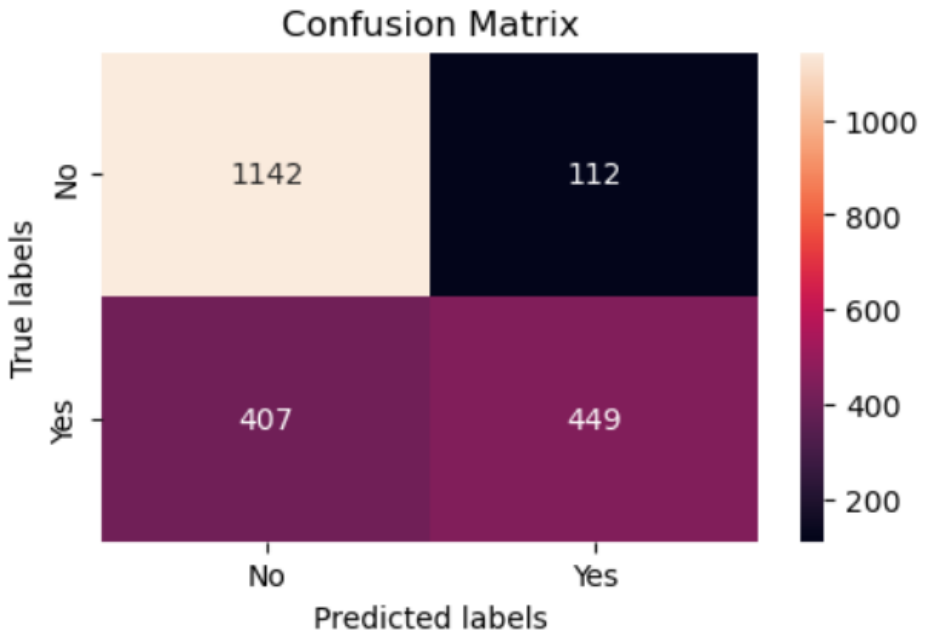


Figure 11: Confusion matrix from the final model

Classification report from the final model

	precision	recall	f1-score	support
No	0.74	0.91	0.81	1254
Yes	0.80	0.52	0.63	856
accuracy			0.75	2110
macro avg	0.77	0.72	0.72	2110
weighted avg	0.76	0.75	0.74	2110

Conclusion and Recommendations

In summary, the modeling results indicate that the prediction accuracy for overall churn behavior is 75%, slightly below the 80% target. The model excels in predicting customers who did not churn (91% recall) but performs less effectively for those who did churn (52% recall). Ongoing improvement is expected with the accumulation of more data, leading to enhanced predictive capabilities over time.

Additional Insights:

Tenure: The churn rate diminishes with increasing tenure, suggesting that longer-tenured customers are less likely to churn. This trend implies that customer loyalty tends to strengthen over time, reflecting higher satisfaction and a more robust relationship with the company.

Payment Method: A correlation exists between the customer's payment method and churn rate. Customers using electronic or mailed checks experience higher churn rates, while those using bank transfer or credit cards with automatic payments exhibit lower churn rates. This indicates that customers opting for more automated and convenient payment methods may have higher satisfaction and loyalty.

Paperless Billing: Customers embracing paperless billing show a higher churn rate compared to those favoring traditional billing methods. This suggests that customers adopting digital processes may have distinct expectations or experiences influencing their decision to churn.

Recommendations:

Focus on New Customer Retention: Implement strategies to enhance satisfaction and engagement for new customers, recognizing that longer tenures correlate with lower churn rates. Provide personalized onboarding experiences, exceptional customer support, and incentives to foster loyalty.

Incentivize Automated Payment Methods: Encourage the use of bank transfers or credit cards with automatic payments by offering discounts or rewards. This promotes customer convenience and loyalty.

Improve Paperless Billing Experience: Identify and address concerns related to paperless billing. Enhance the user interface, communicate billing details clearly, and offer additional benefits to customers opting for paperless billing.

Conduct Customer Satisfaction Surveys: Regularly collect feedback to understand customer needs, preferences, and satisfaction levels. This proactive approach can help identify areas for improvement and address customer concerns.

Implement Targeted Marketing Campaigns: Focus on educating customers about the company's services, addressing common pain points, and highlighting the advantages of long-term relationships.

Continuous Monitoring and Model Updates: Regularly monitor and update the churn prediction model with new data. Customer churn patterns evolve, and keeping the model accurate ensures effective predictions over time.