Customer Churn Modelling for Internet Service Providers

by Nabila Ijtihatil Ismail

Table of Contents

Business Problem

Due to intense competition and the increasing variety of services offered by many providers, customer churn is a common challenge for ISPs. The cost of acquiring new customers is higher than that of retaining current ones. Therefore, customer retention becomes a critical priority.

Objectives

1. What are the habits of customers who churn and those who don’t churn?
2. How can we build effective machine learning models to predict customer churn?

Data Understanding

#	Column	Non-Null Count	Dtype
0	id	72274 non-null	int64
1	is_tv_subscriber	72274 non-null	int64
2	is_movie_package_subscriber	72274 non-null	int64
3	subscription_age	72274 non-null	float64
4	bill_avg	72274 non-null	int64
5	reamining_contract	50702 non-null	float64
6	service_failure_count	72274 non-null	int64
7	download_avg	71893 non-null	float64
8	upload_avg	71893 non-null	float64
9	download_over_limit	72274 non-null	int64
10	churn	72274 non-null	int64

see the dataset **here**!

- The dataset contain basic information for internet service subscriber
- This dataset has 11 columns and 72.274 rows
- There are **21.572 missing values in the reamining_contract** column, which is about 29.8% of the total data.
- There are **381 missing value in the download_avg and upload_avg** column, which is about 0.5% of the total data
- No duplicated values

Data Preprocessing

01 Handling Missing Value

- The missing values in the remaining_contract column can be replaced with 0, as the blank entries indicate that the customer does not have an active contract.
- The missing values in download_avg and upload_avg can be filled with the median values because the data has outlier.

02 Handling Abnormal Data

There are some abnormal data entries, such as subscription_age being less than 0 or negative. These values can be corrected by replacing them with the median value because the data has outlier.

03 Handling Outliers

No remove, the outlier is reasonable

Data Preprocessing

04 Rename Column

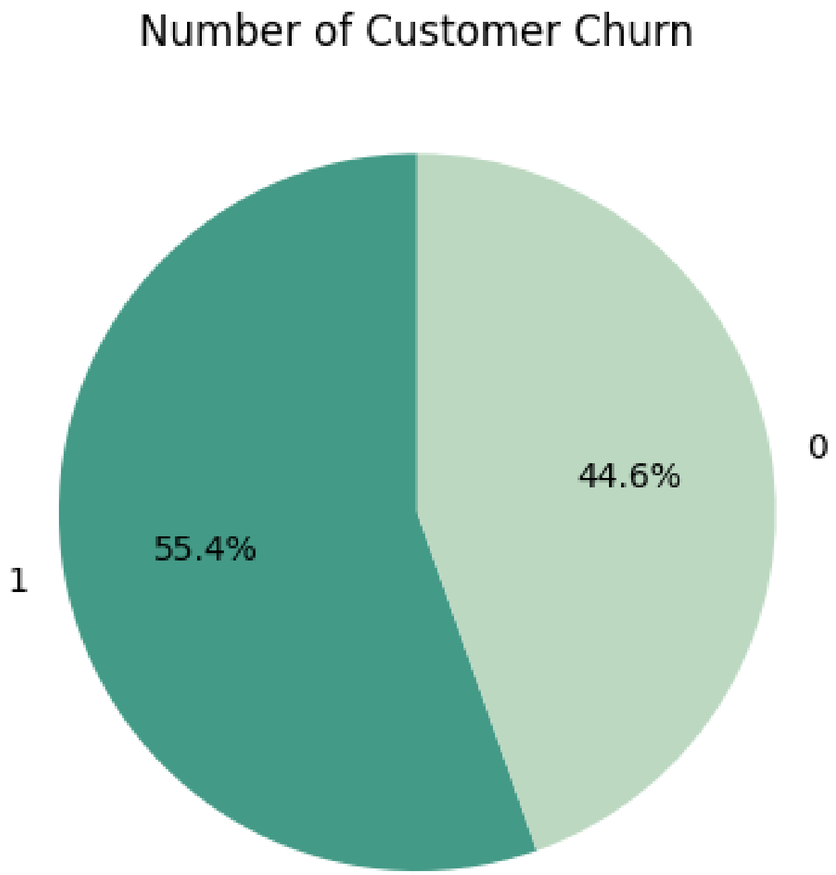
Rename columns for better understanding, such as changing 'id' to 'customer_id' and 'reamining_contract' to 'remaining_contract'.

05 Add New Column

create a new column 'subscribe' with the following conditions:

1. is_tv_subscriber = 0 and is_movie_package_subscriber = 0 then internet only
2. is_tv_subscriber = 1 and is_movie_package_subscriber = 0 then tv only
3. is_tv_subscriber = 0 and is_movie_package_subscriber = 1 then movie only
4. is_tv_subscriber = 1 and is_movie_package_subscriber = 1 then both

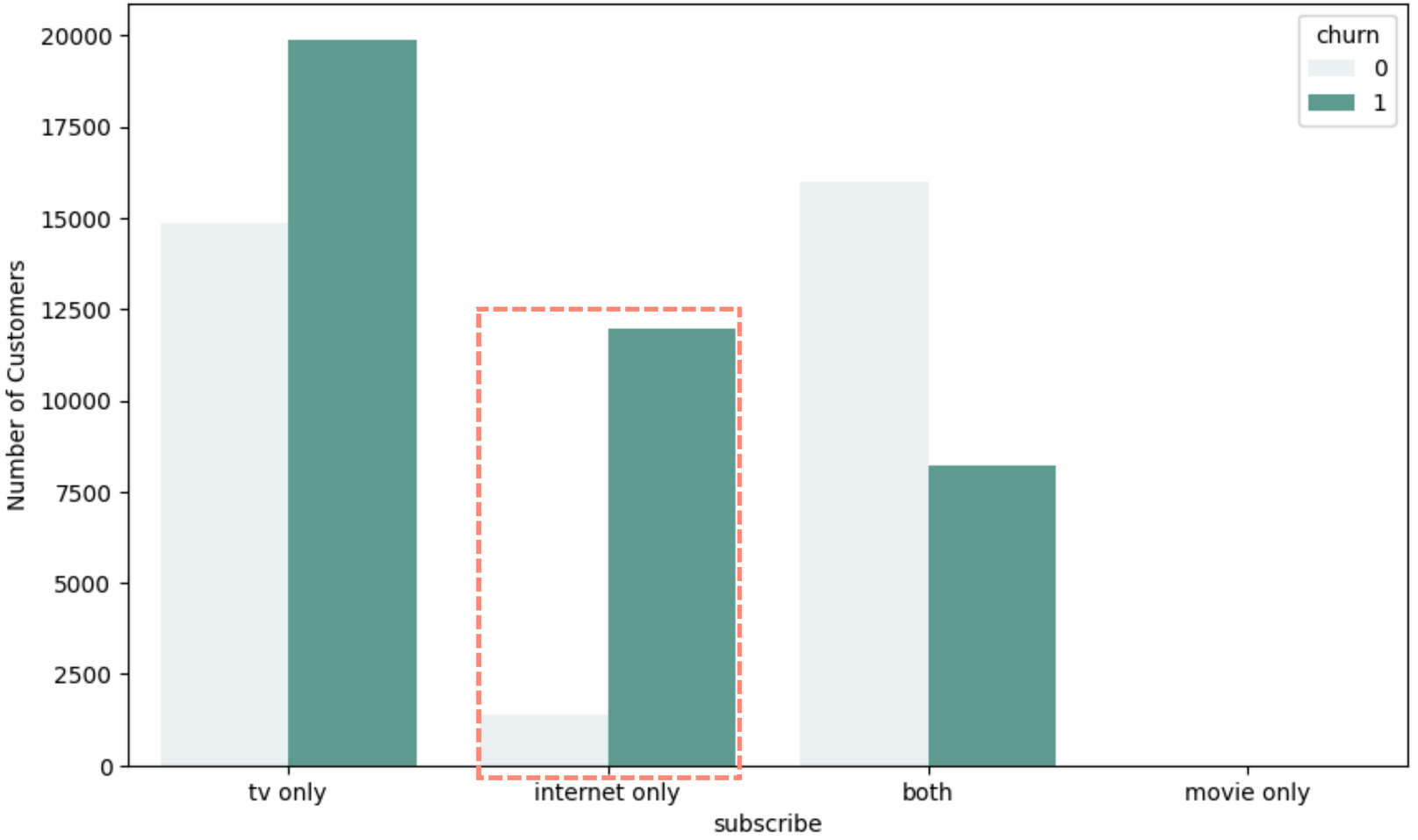
Exploratory Data Analysis



How many customer churn?

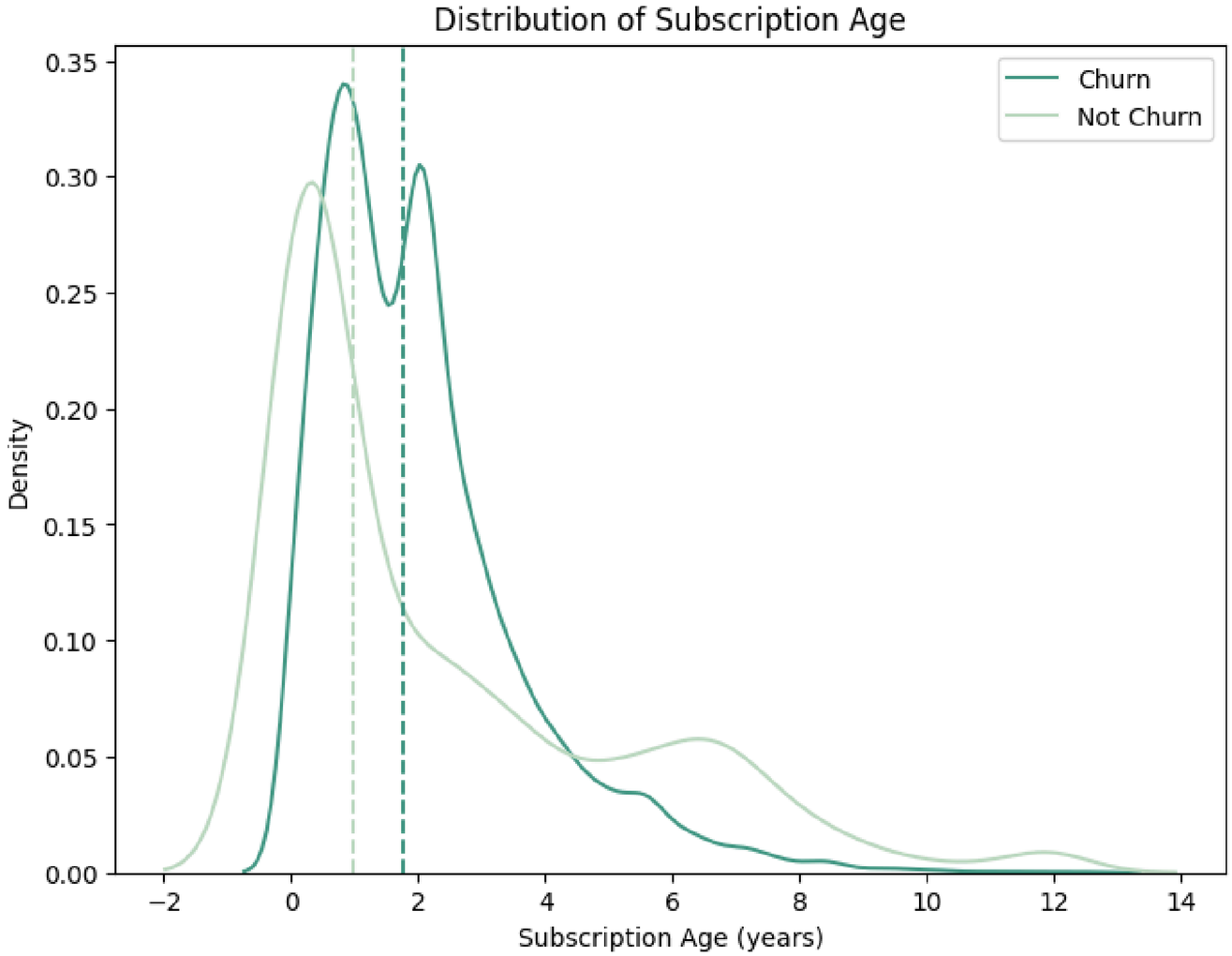
There are **more than 50% of customers who have churned**, which is around 40,050 people have already cancelled the service.

How about their subscription packages choices?



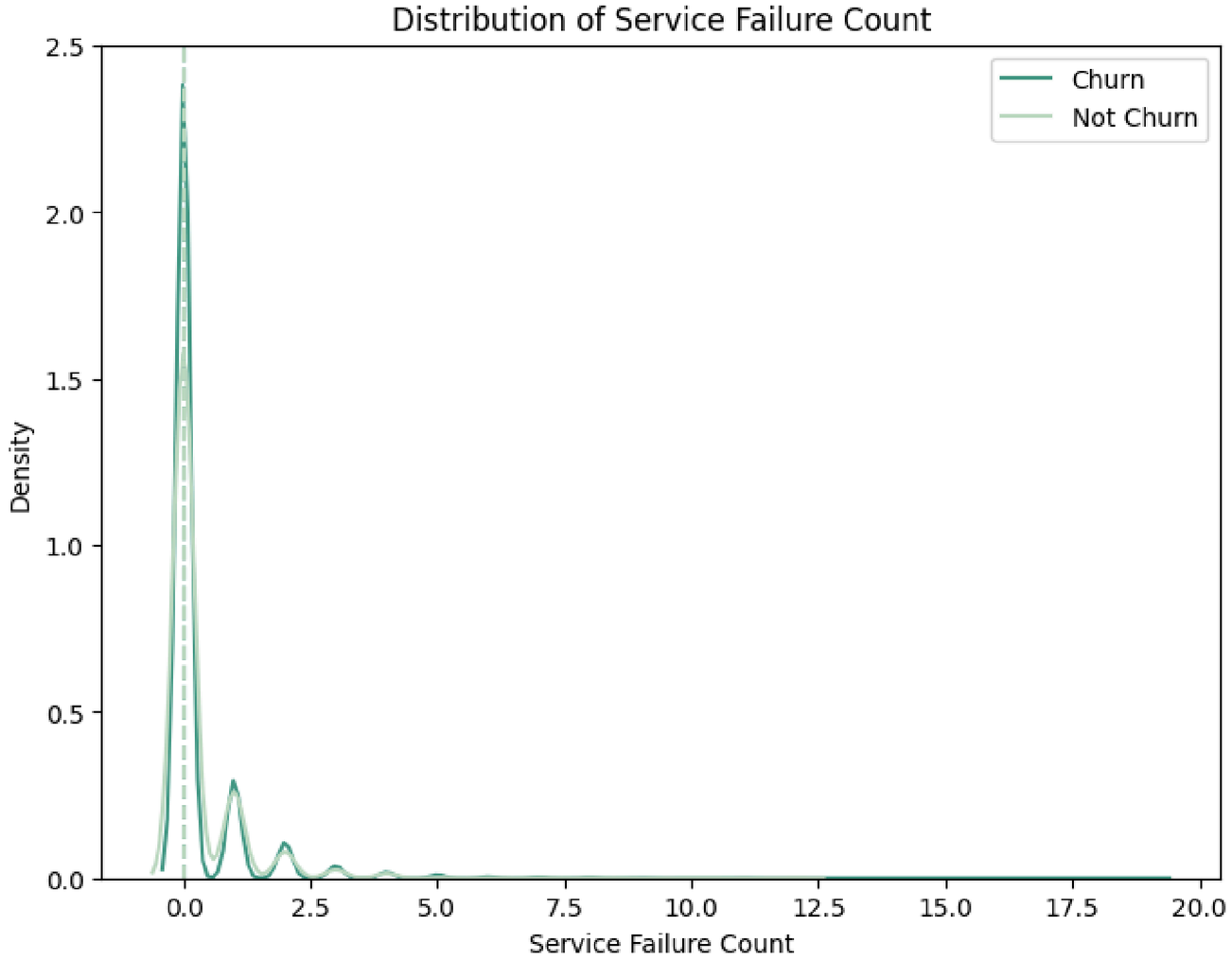
Customers who subscribe to internet only has higher churn rate. This suggests that customers who **canceled the service mostly came from customers who subscribed to the internet only.**

How is the distribution of the subscription age?



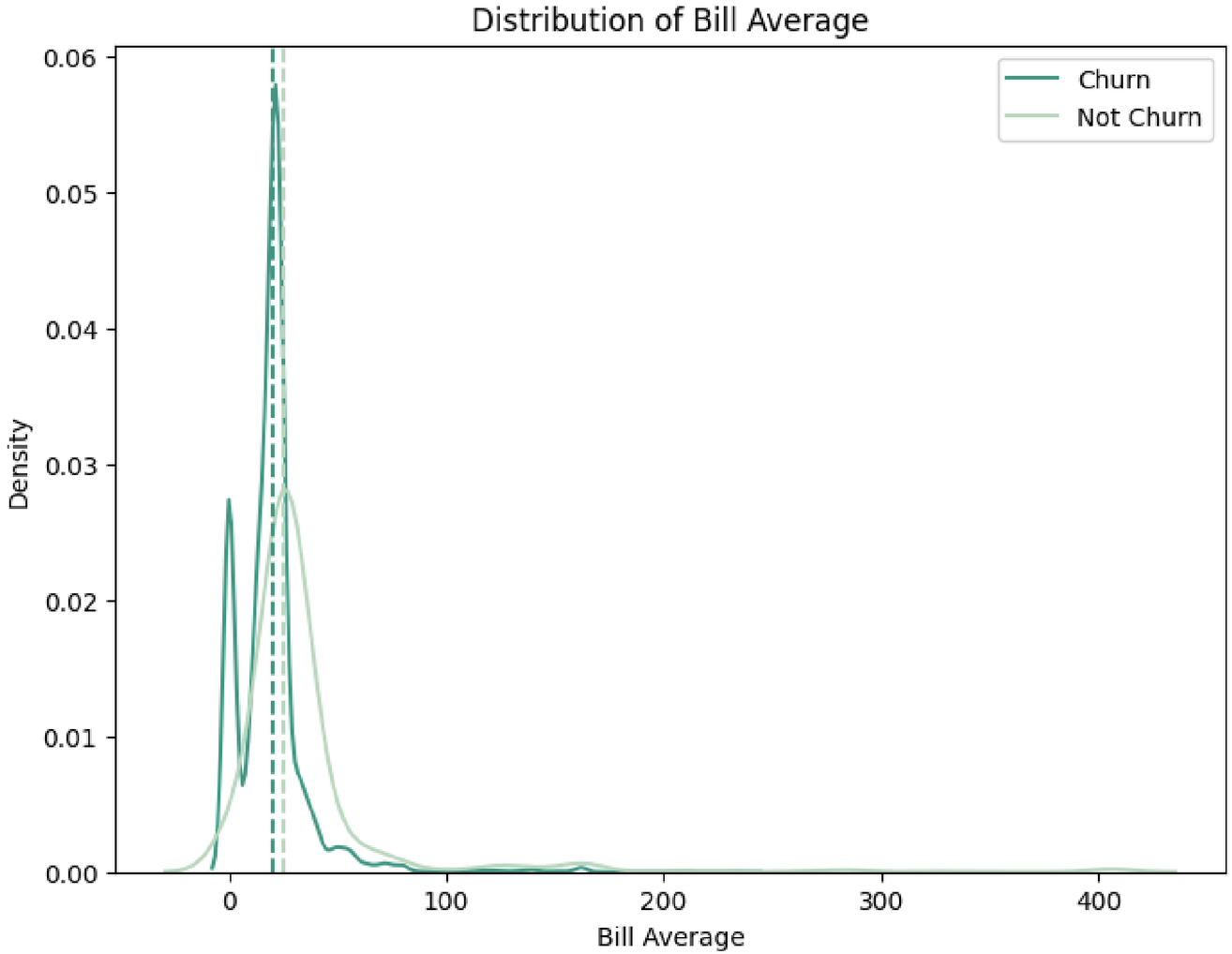
Churned customers usually have shorter subscription ages than non-churned. This suggests that churned customers canceled the service within the first few years of their subscription.

How many service failures did they experience?



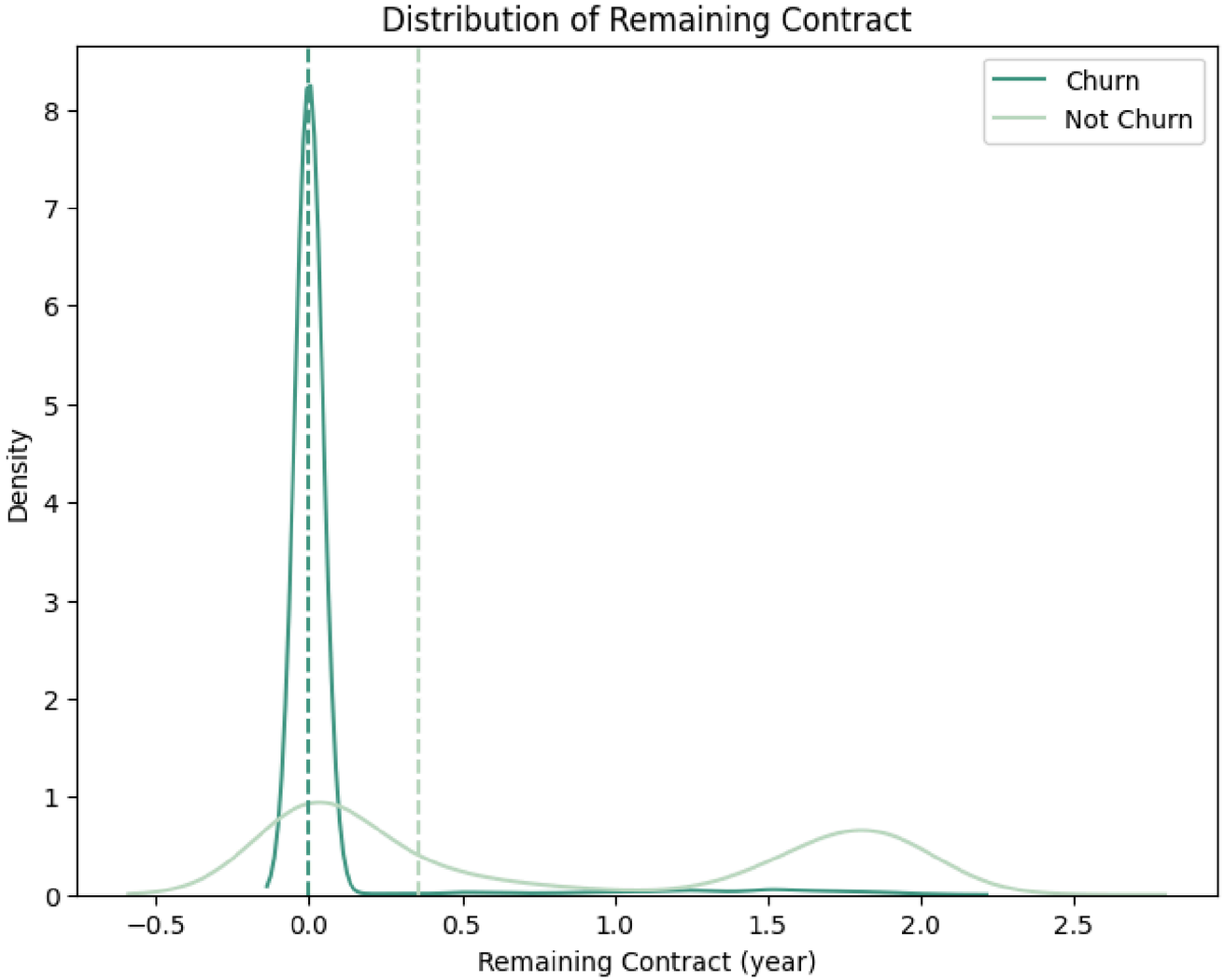
Churned customers tend to experience more service failures than non-churn customers. This indicates that service failures may contribute to customer dissatisfaction and churn.

What is the total billing amount they receive?



Churned customers tend to have slightly lower bill averages than those who do not churn. This may indicate that churned customers may still feel that the cost is too high for them or may not see any added value for them to continue subscribing.

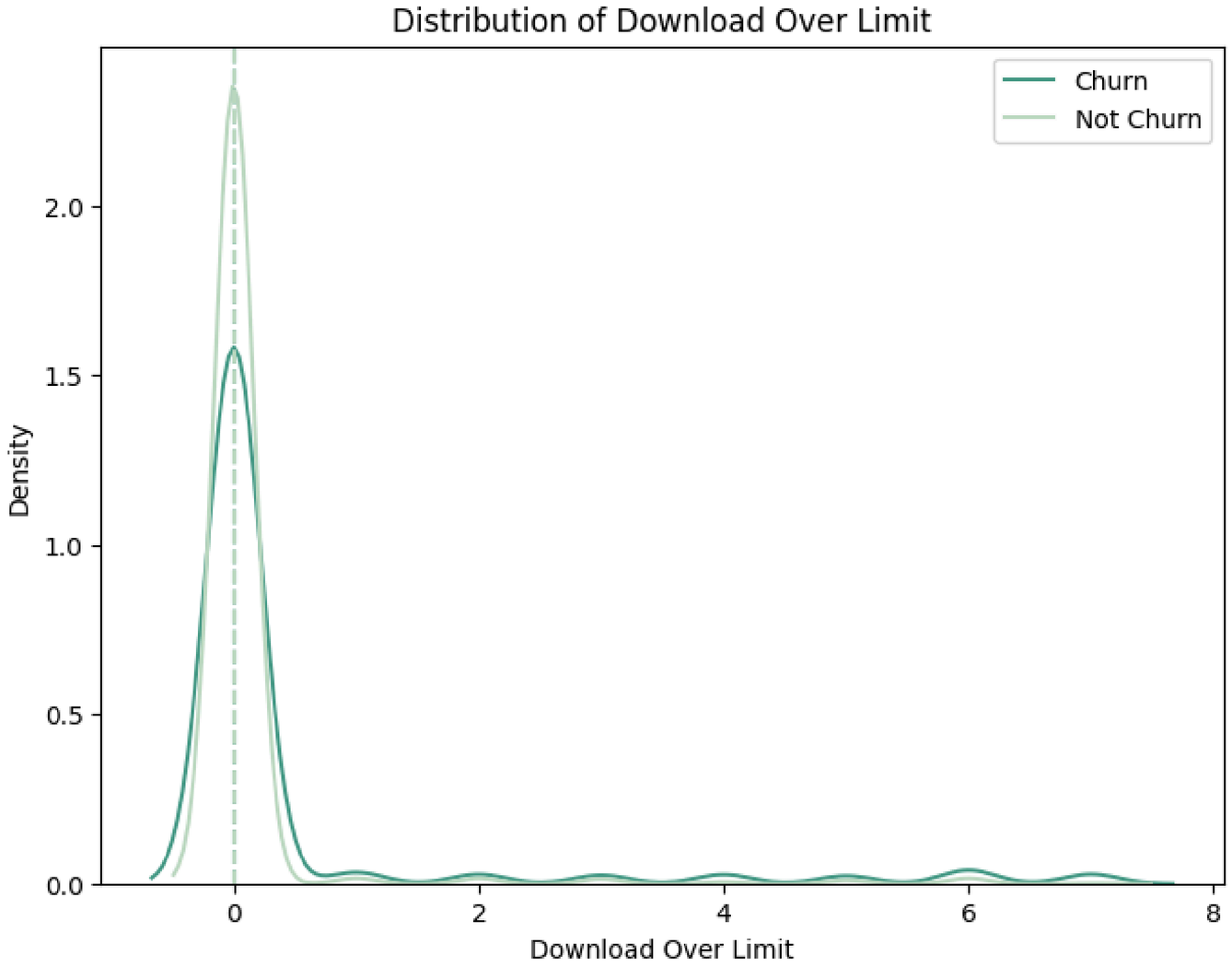
How many contracts remain active?



Churned customers mostly do not have an active contract.

But there are some customers who had contracts but still churned. This may suggest dissatisfaction with service despite the contractual penalty.

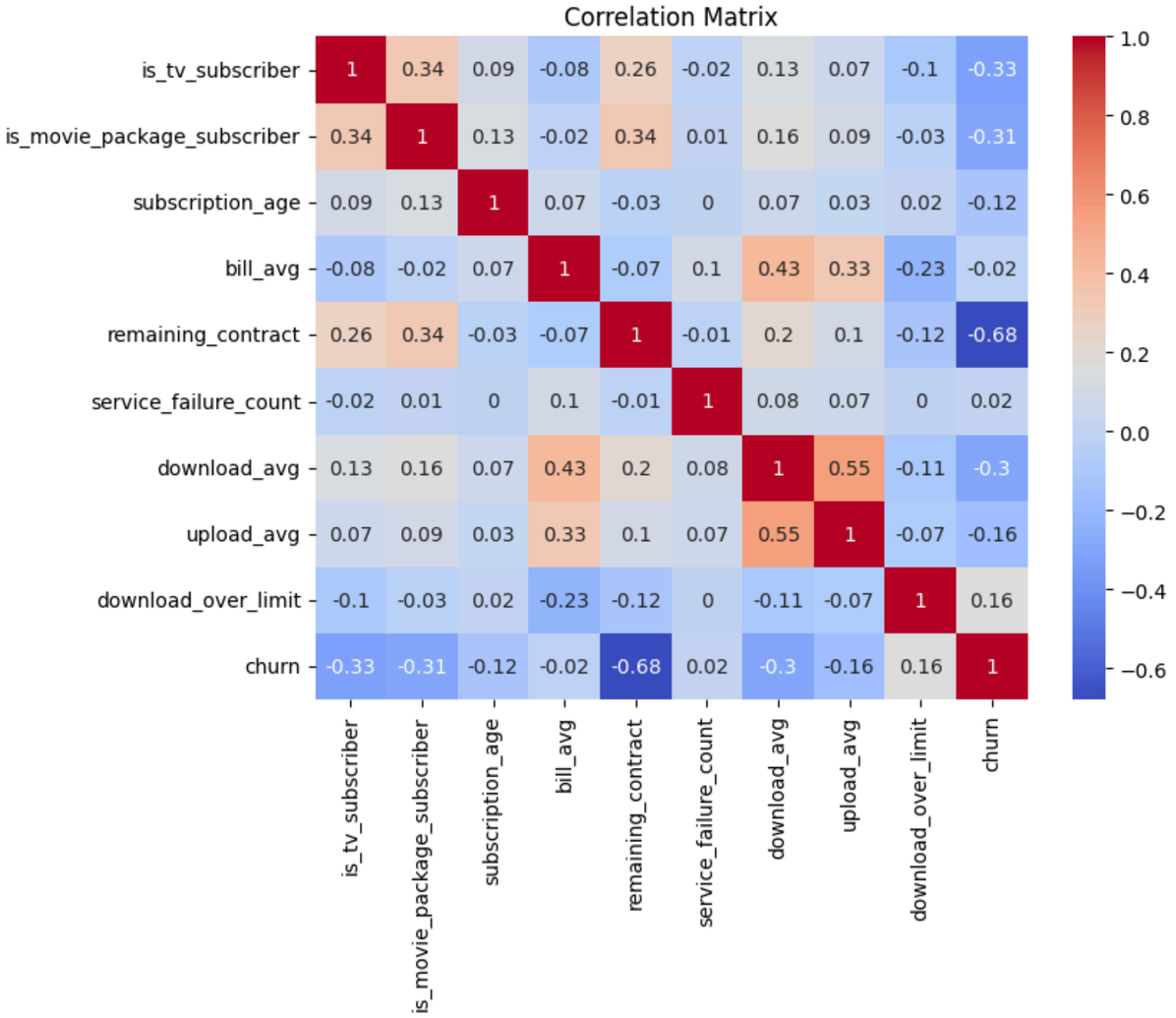
Do any customers exceed their data usage limits?



Churned customers tend to exceed their data limits more frequently than non-churned customers so they are charged extra, which causes dissatisfaction.

Correlation Matrix

The purpose of using a correlation matrix is to avoid multicollinearity or high correlation between variables. Based on the results, **no variables have a high correlation** so that all variables can be used.



Feature Scaling

- encode ‘subscribe’ column using LabelEncoder()
- scale numerical variables using RobustScaler()

Machine Learning Models

The data is divided into two parts:
80% for the training set and 20% for the testing set.

Comparing Models

	model	recall_train	recall_test
0	Logistic Regression	91.834379	91.863517
1	Decision Tree	99.996880	91.301087
2	Random Forest	100.000000	93.713286
3	LightGBM	94.355518	93.825772
4	XGBoost	95.375831	94.013248

Good Fit Models : Logistic Regression, LightGBM, XGBoost

Overfitting Models: Decision Tree, Random Forest

Recall: Minimize the occurrence of false negative prediction errors.

Conclusion

1. The majority of customers who churn are those subscribed only to internet services (without additional tv or movie services)
2. Churned customers mostly canceled the service within the first 1-2 years of subscription age
3. Churned customers tend to experience more service failures
4. Churned customers' average bill tends to be lower, around $25.
5. Most churned customers do not have an active contract, though there are some who do and still decide to leave.
6. Churned customers are more likely to exceed their data limits more than twice, leading to additional charges.
7. The churn prediction model will help identify customers likely to cancel their services.

Recommendation

1. Offer a **10% discount on the next bill** for customers who agree to extend their subscription
2. Offer a **promotional package** that includes **additional TV channels and movie packages** for a limited time.
3. Optimize model performance through hyperparameter tuning to prevent overfitting and other feature engineering

Data Analyst, Business Intelligence

Thank you!
reach me through:

nabila.ijtihatil@gmail.com

www.linkedin.com/in/nabila-ijtihatil-ismail/

github.com/nabilaijtihatil