Let's Start

Telco Customer Churn

By Smart Python

DigitalSkola

Mentor

DigitalSkola

Members of Smart Python

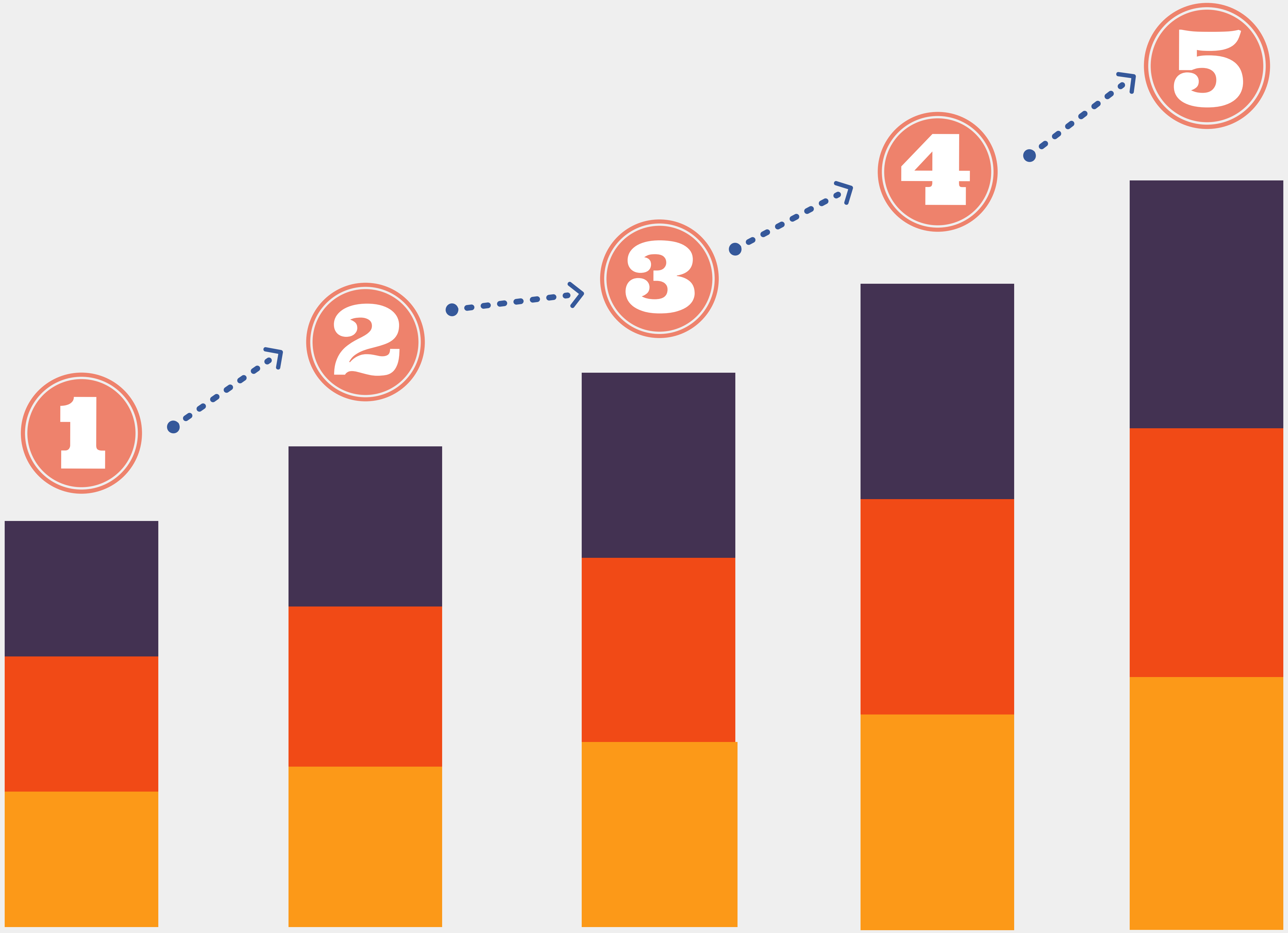
Jaelani

Vanadhia Amanita

Pingki Vila

What We Did

1. Overview of the Business Understanding
2. Identify The Dataset
3. EDA
4. Data Pre-Processing
5. Develop Model and Evaluation
6. Recommendation



Overview of the Business Understanding

1

Telco Company

Telco Company has provided home phone and Internet services to 7043 customers in California in Q3.

Telco Company has a problem like Customer Churn. It indicates which customers have left, stayed, or signed up for their service.

Source: community.ibm.com

Background

- Customer Churn means losing customers from a business.
- Churn is calculated by how many customers leave your business in a time.
- Customer Churn is important for businesses because it is a picture of the success business in retaining customers.

"Predict behavior to retain customers"

Background

Annual churn rates for telecommunications companies average between 10% and 67%.

Industry retention surveys have shown that while price and product are important, most people leave any service because of dissatisfaction with the way they are treated.

It costs hundreds of dollars to acquire a new customer in most Telecom industries.

Source: Database Marketing Institute, 2022

Objectives

1. To identify which the best model in predicting the customer churn
2. To identify which variables are significantly affect the customer churn

2 Identify The Dataset

About Dataset

Churn

Customers who left within the last month

Demographic

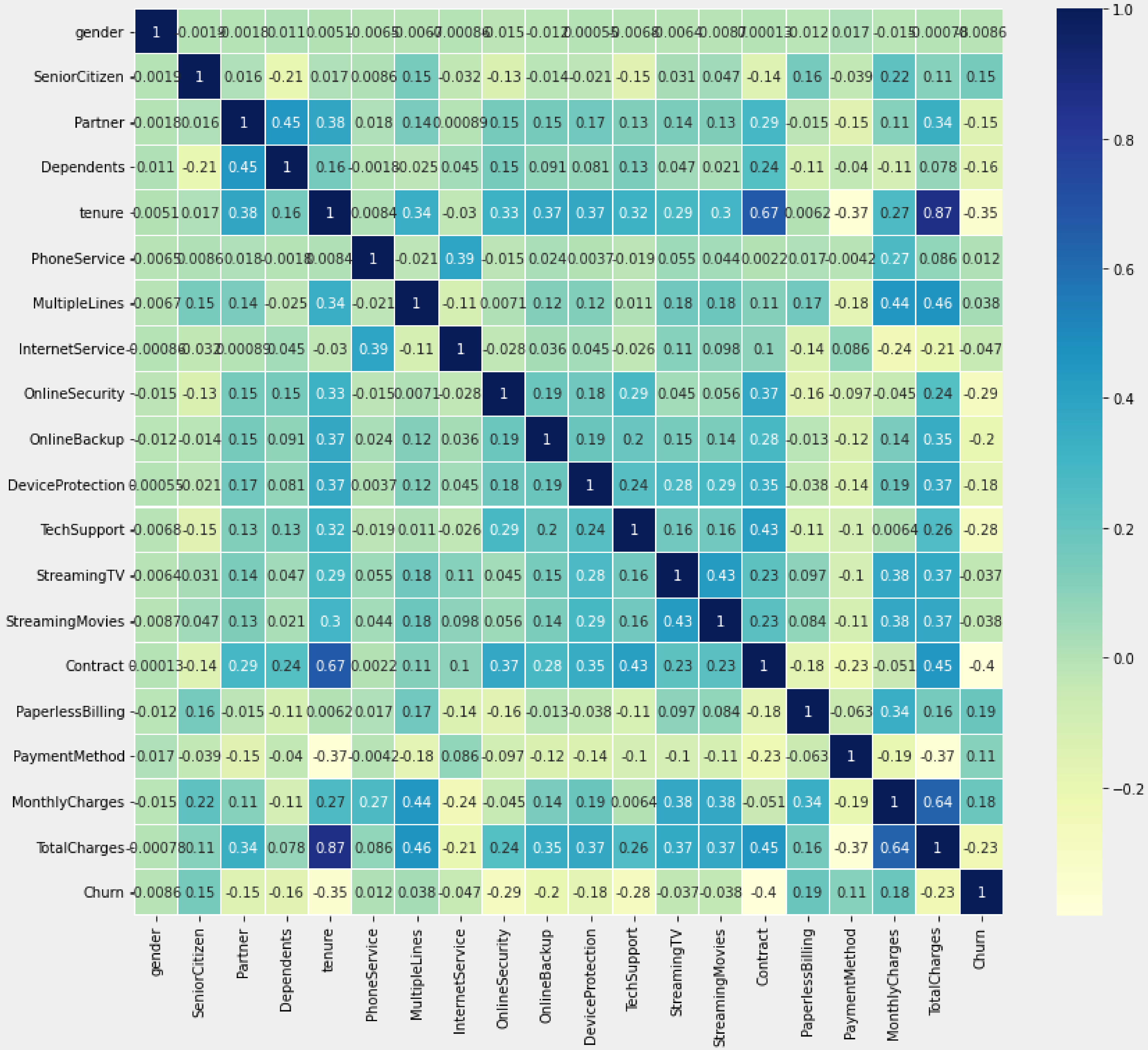
Gender
Senior Citizen
Partners
Dependents

Account Information

Tenure
Contract
Payment Method
Paperless Billing
Monthly Charges
Total Charges

Services

Phone
Multiple lines
Internet
Online security
Online backup
Device protection
Tech support
Streaming TV
Streaming movies



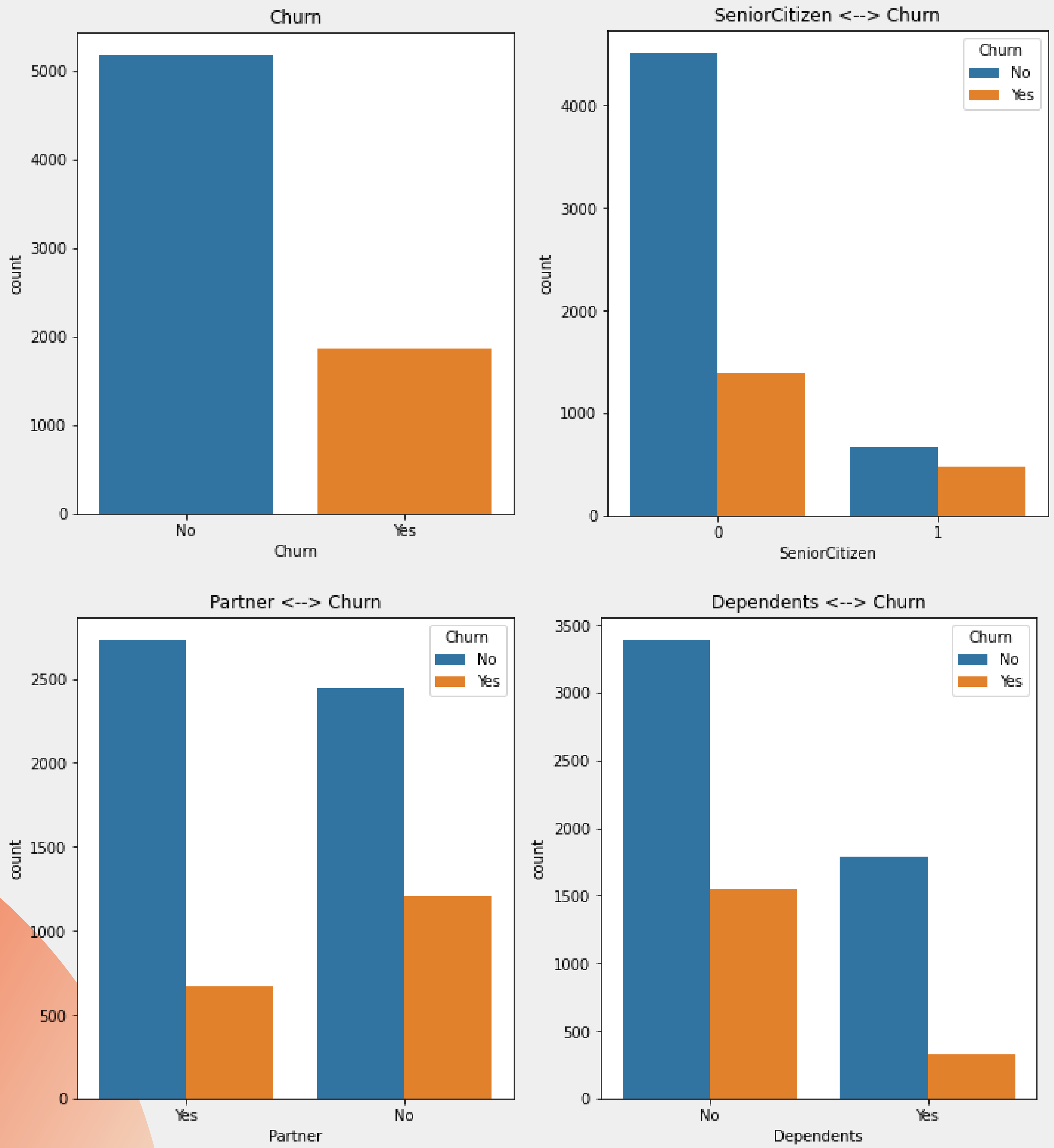
Correlation

- TotalCharges and tenure have strong positive correlation
- Tenure and contracts have a strong correlation
- Monthly charges and total charges have strong positive correlation

3

Exploratory Data Analysis

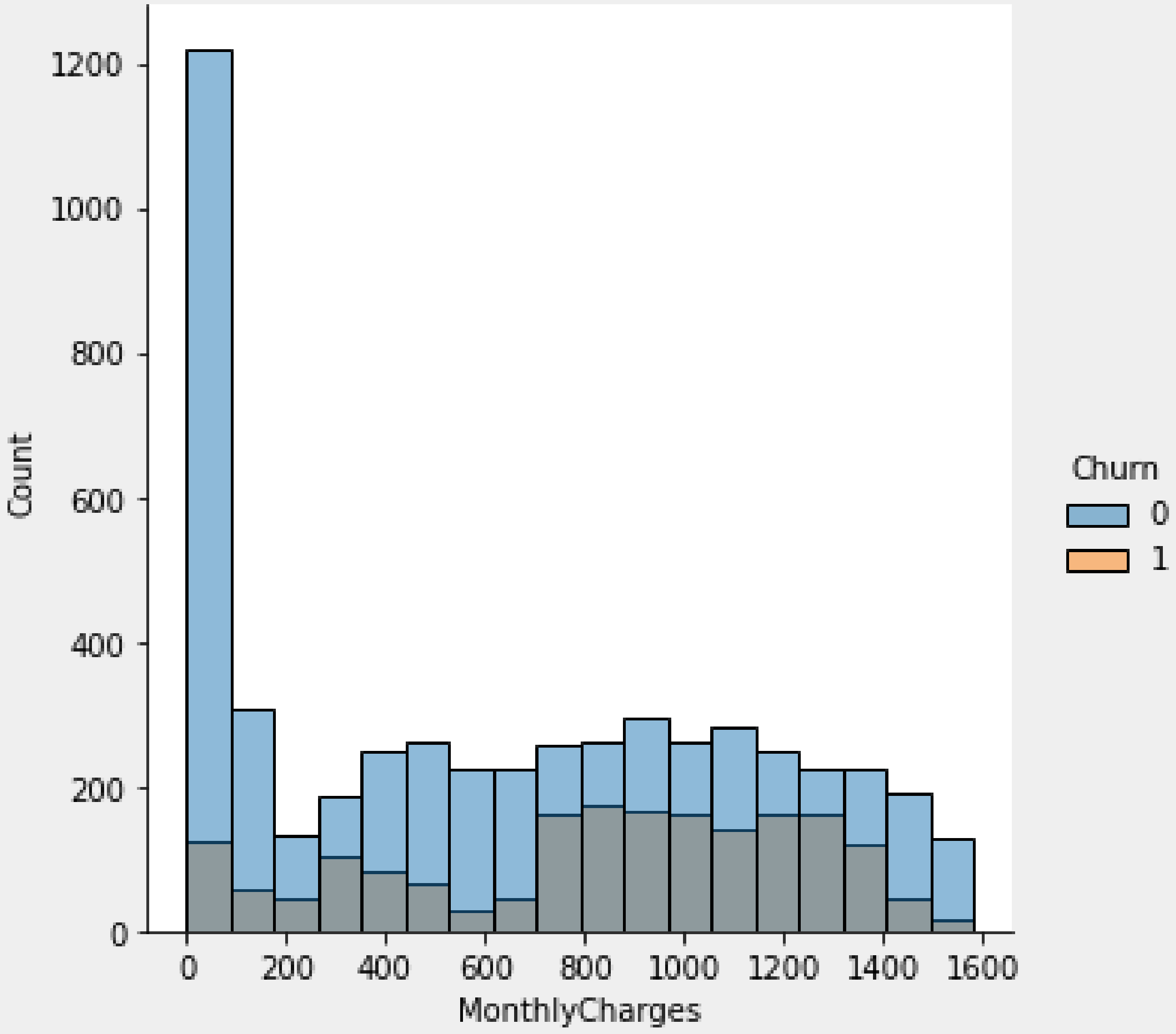
Demographic data

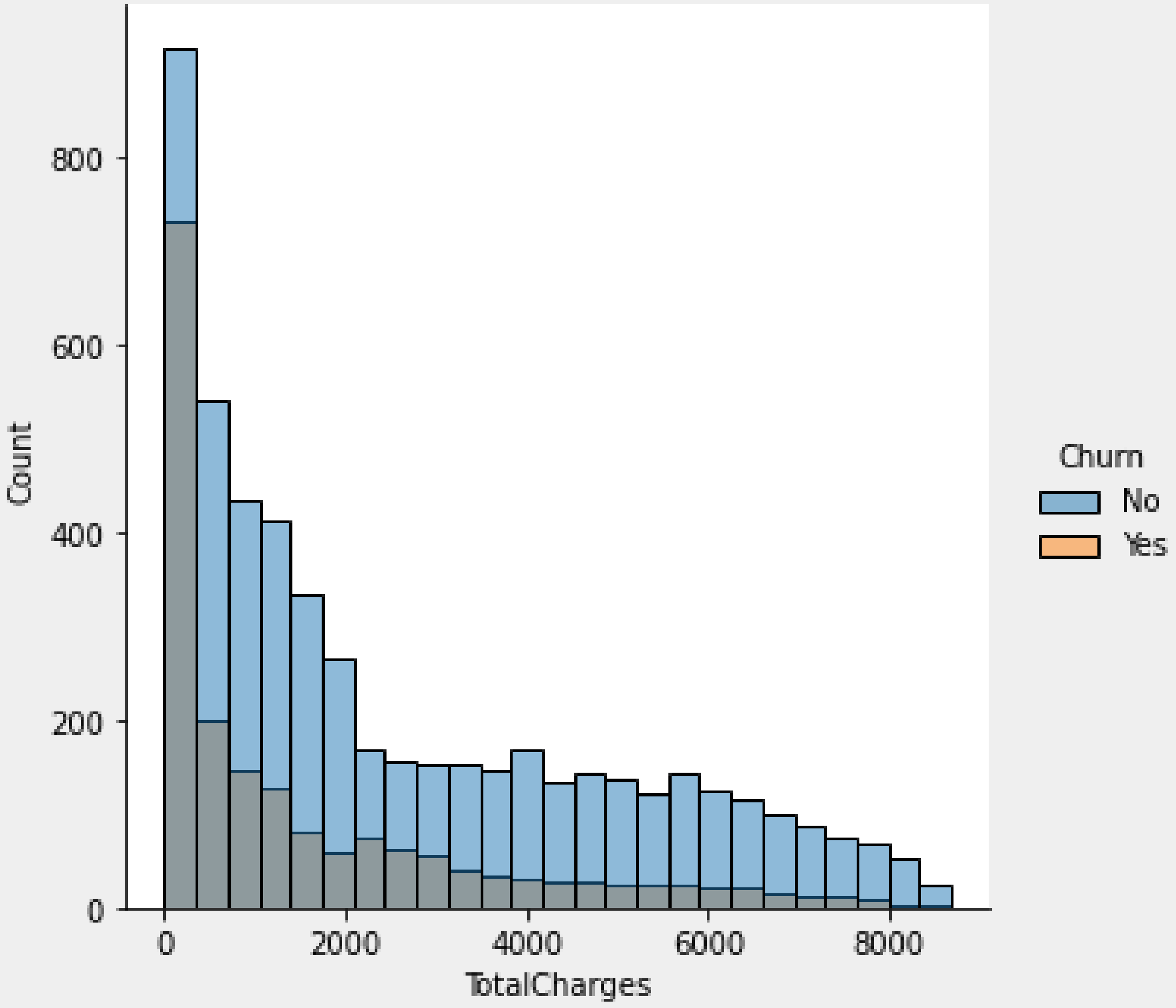


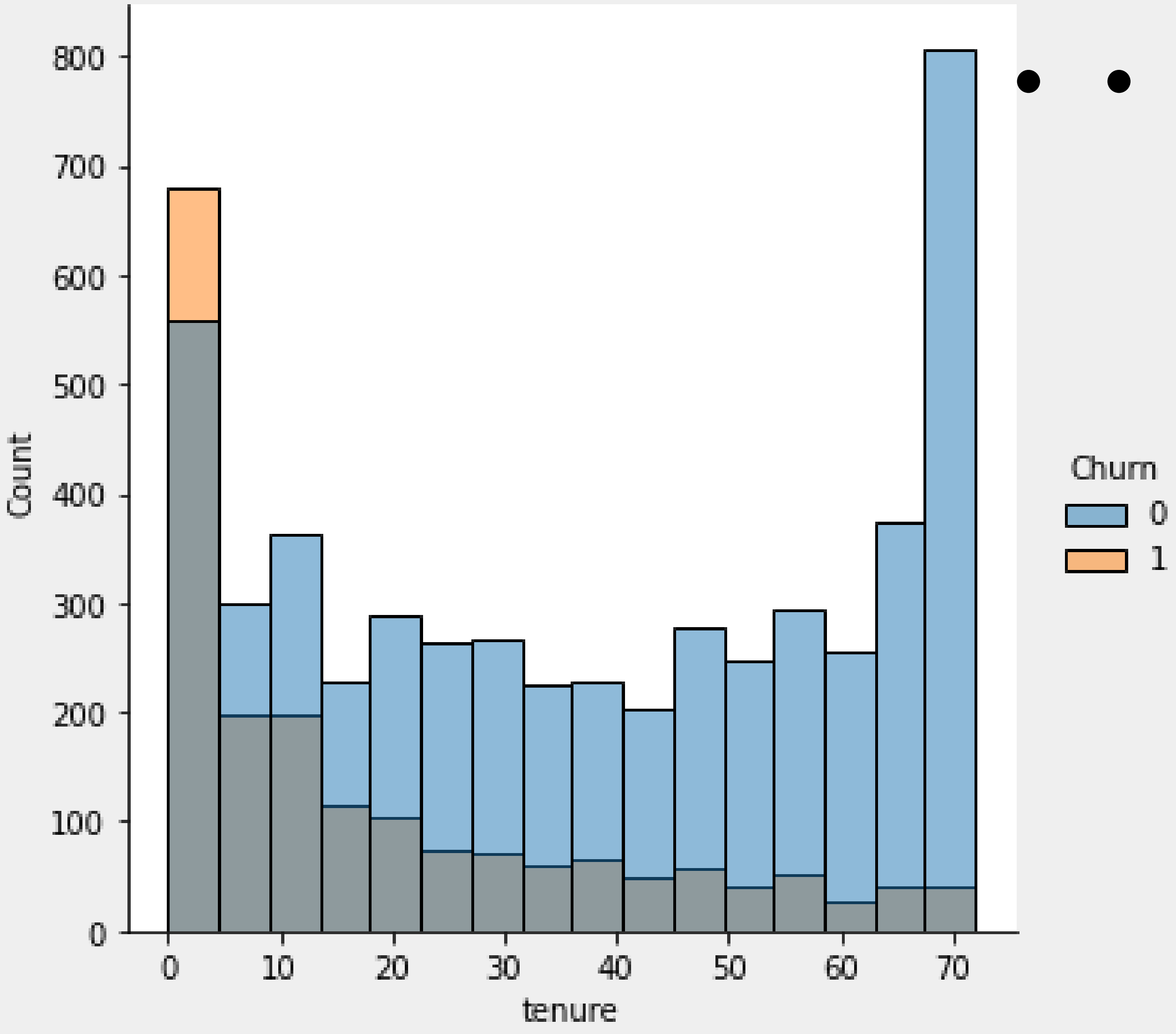
According to the charts, the following are characteristics of customers that are churn:

1. Customer with no partner
2. Customer with no dependents
3. Customer who are younger

Account Information







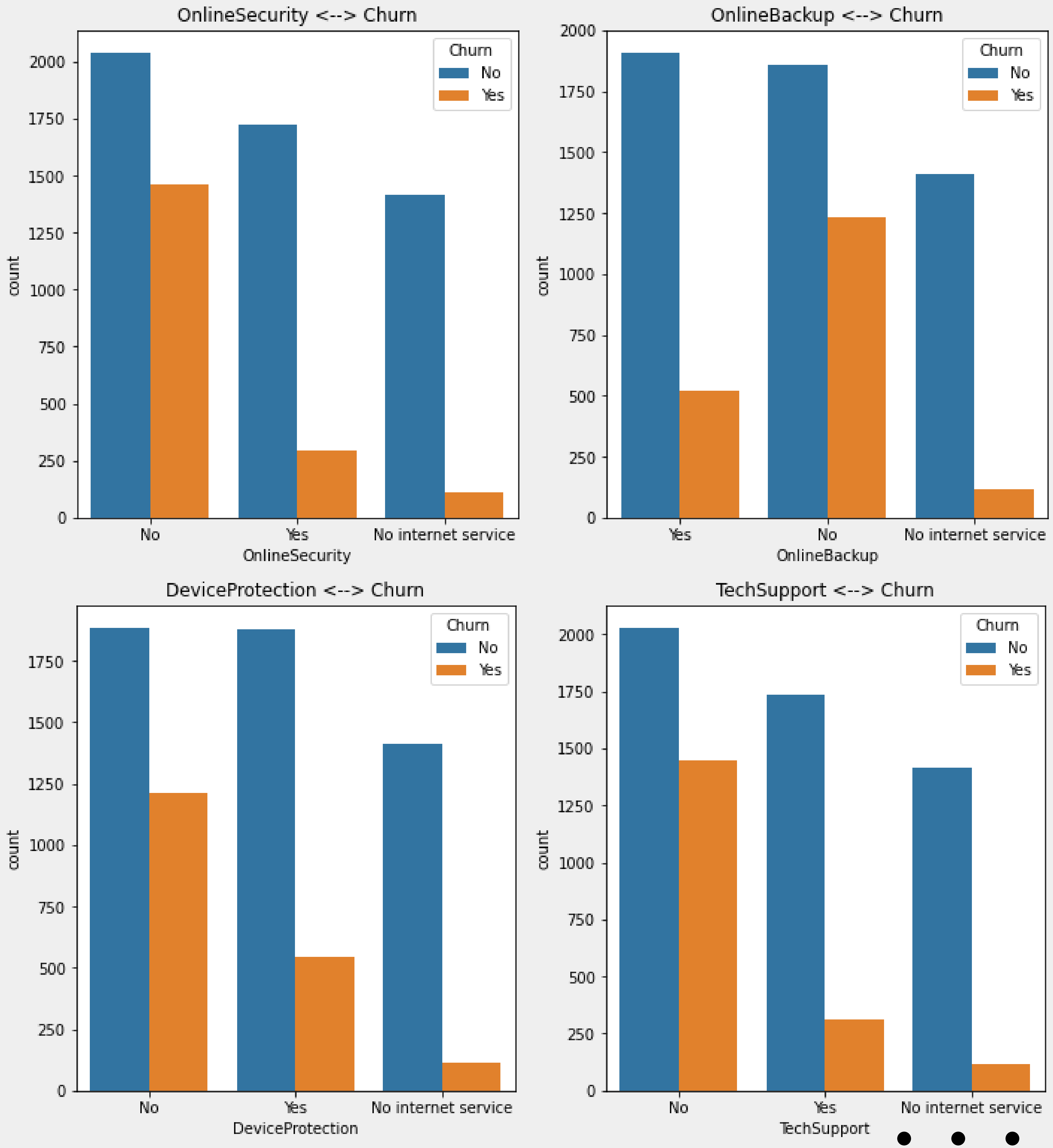
Customer with higher monthly charges are likely to churn, affecting the tenure to be shorter

Services

- Customers are likely to churn when they don't have online security, online backup, device protection, and tech support.

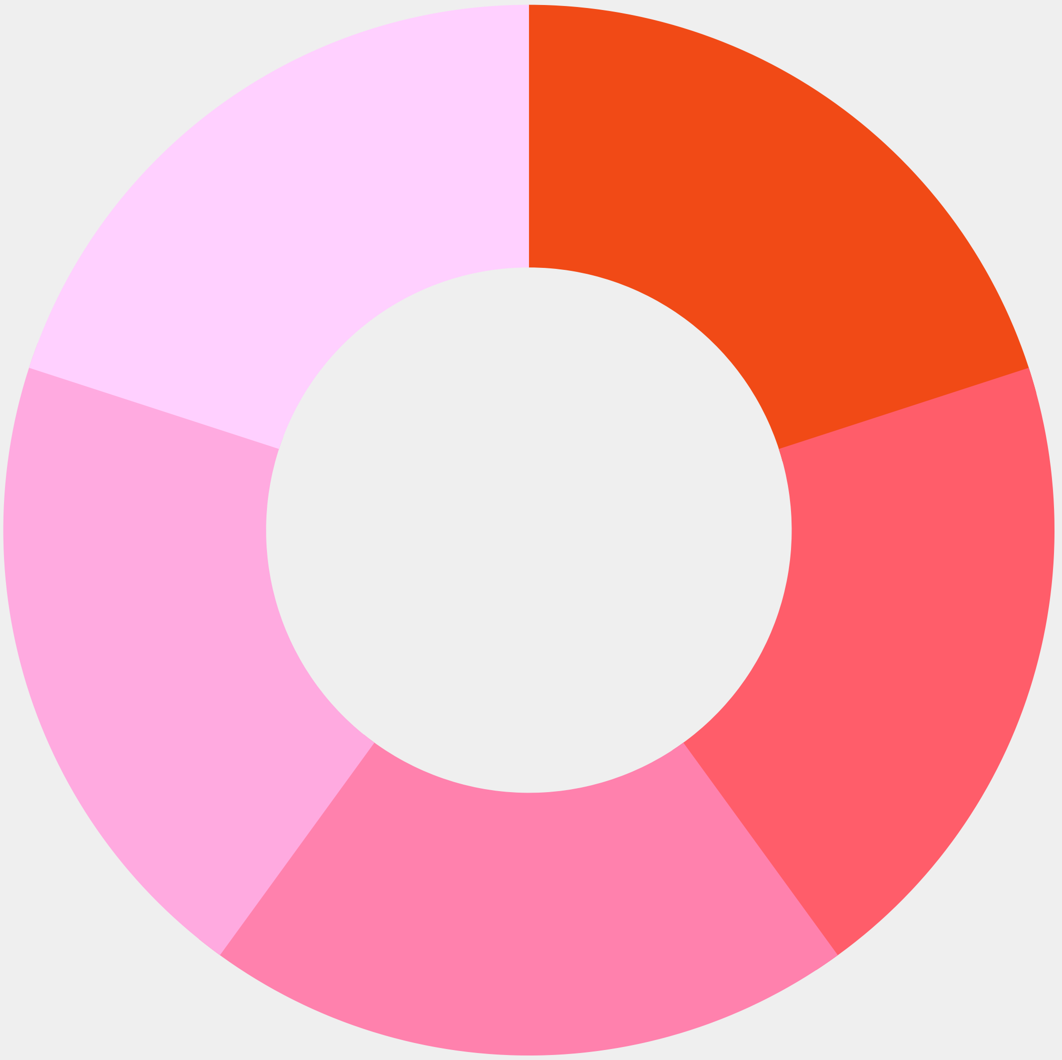
- Mostly because of small price differences of subscription to these service and without service

		count	mean
		MonthlyCharges	MonthlyCharges
Churn	OnlineSecurity		
No	No	2037	74.625233
	No internet service	1413	21.136058
	Yes	1724	78.369432
Yes	No	1461	77.181896
	No internet service	113	20.368142
	Yes	295	81.581356

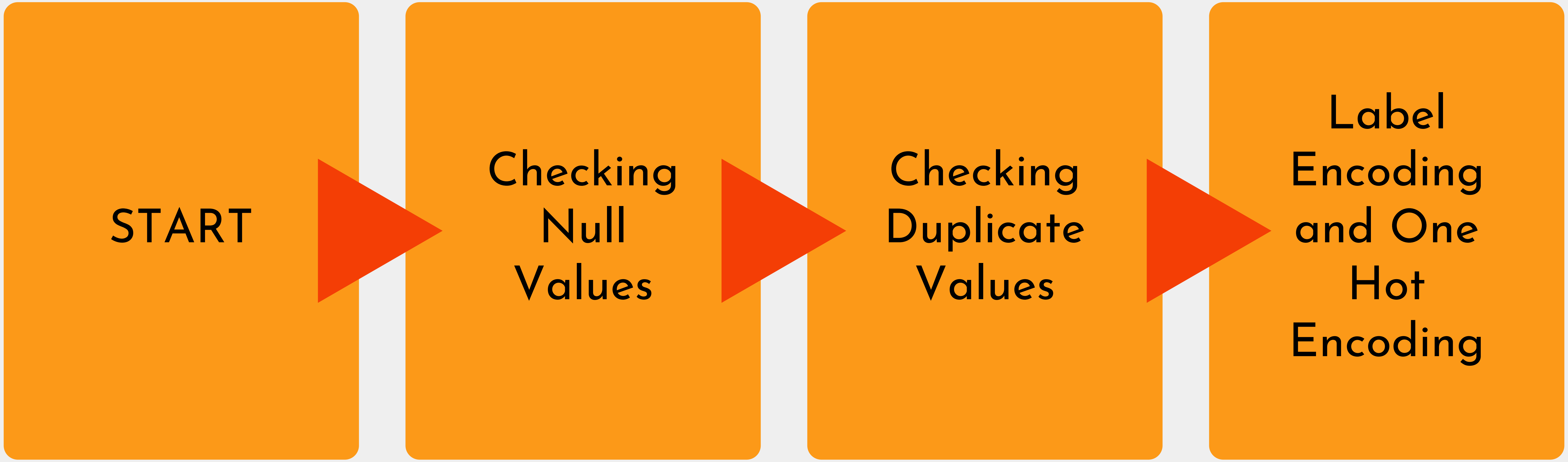


4

Data Pre-Processing



Steps on Data Pre-Processing



Checking Null Values

```
#Check missing values
data.isnull().sum()
```

```
customerID          0
gender              0
SeniorCitizen       0
Partner             0
Dependents          0
tenure              0
PhoneService        0
MultipleLines       0
InternetService     0
OnlineSecurity      0
OnlineBackup        0
DeviceProtection    0
TechSupport         0
StreamingTV         0
StreamingMovies     0
Contract            0
PaperlessBilling    0
PaymentMethod       0
MonthlyCharges      0
TotalCharges        0
Churn               0
dtype: int64
```

Checking Duplicate Values

```
# check duplicate data
data.duplicated().sum()
```

```
0
```

Feature Encoding

One-Hot Encoding

Change each category so that it has a value of 1 or 0

```
#one hot encode
df_onehot = pd.get_dummies(df_X, columns=['gender', 'Partner', 'Dependents', 'PhoneService', 'MultipleLines',
       'InternetService', 'OnlineSecurity', 'OnlineBackup', 'DeviceProtection',
       'TechSupport', 'StreamingTV', 'StreamingMovies', 'Contract',
       'PaperlessBilling', 'PaymentMethod'], drop_first=False)
df_onehot.head()
```

	gender_Female	gender_Male	Partner_No	Partner_Yes	Dependents_No	Dependents_Yes	PhoneService_No	PhoneService_Yes
0	1	0	0	1	1	0	1	0
1	0	1	1	0	1	0	0	1
2	0	1	1	0	1	0	0	1
3	0	1	1	0	1	0	1	0
4	1	0	1	0	1	0	0	1

5 rows × 41 columns

Feature Encoding

Label Encoding

Converts each category to numbers 1,2,3, … etc

```
#label encoding (categorical encoding)
cats = df_X.select_dtypes(include=['object', 'bool']).columns
cat_features = list(cats.values)
cat_en = LabelEncoder()
for i in cat_features:
  df_X[i] = cat_en.fit_transform(df_X[i])

df_X
```

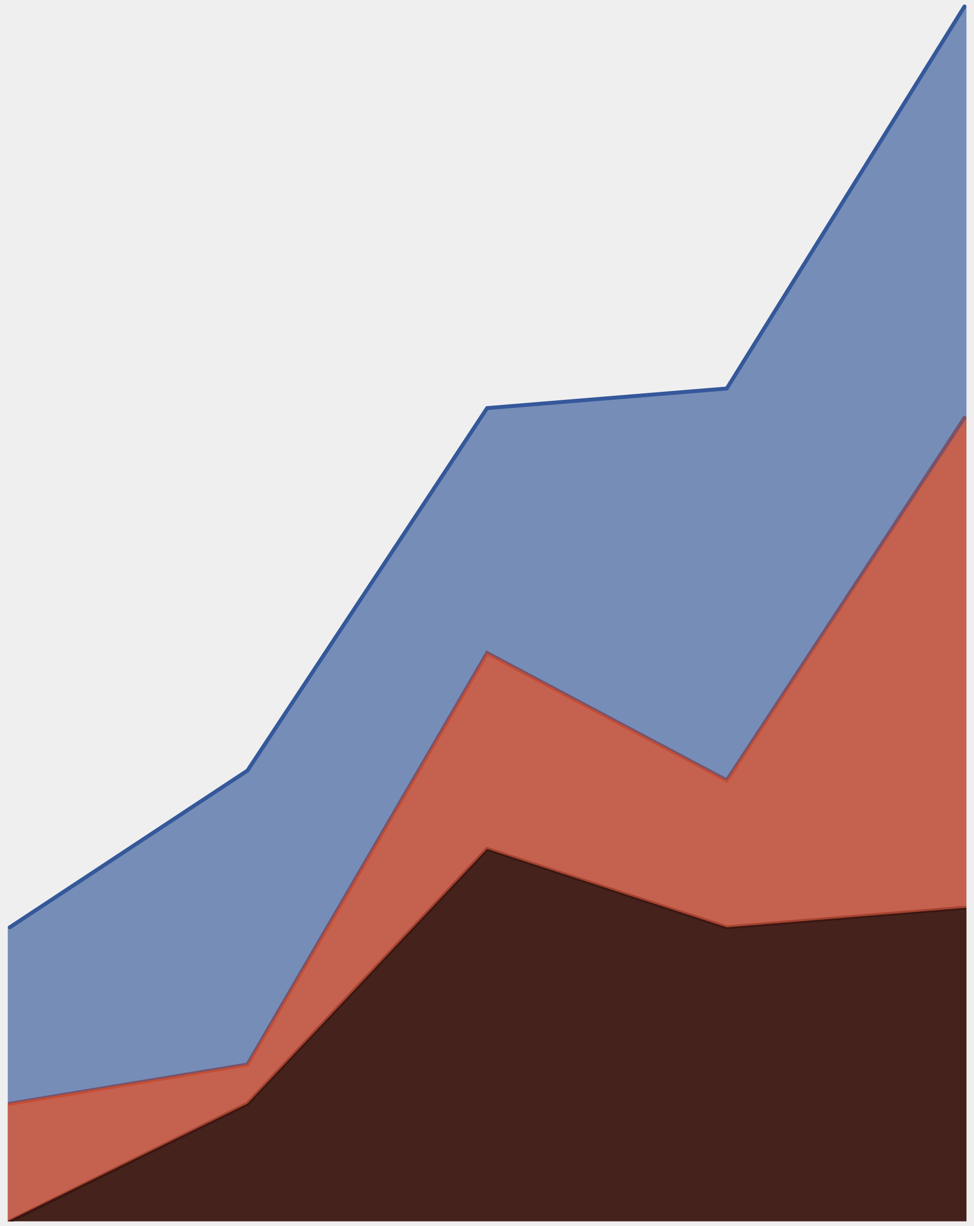
	gender	Partner	Dependents	PhoneService	MultipleLines	InternetService	OnlineSecurity
0	0	1	0	0	1	0	0
1	1	0	0	1	0	0	2
2	1	0	0	1	0	0	2
3	1	0	0	0	1	0	2
4	0	0	0	1	0	1	0
...	...	...	...	...	...	...	...
7038	1	1	1	1	2	0	2
7039	0	1	1	1	2	1	0
7040	0	1	1	0	1	0	2
7041	1	1	0	1	2	1	0
7042	1	0	0	1	0	1	2

```
#label encoding for y
le = LabelEncoder()
le.fit(df_y)
df_y= le.fit_transform(df_y)
df_y
```

```
array([0, 0, 1, ..., 0, 1, 0])
```

5

Develop Model, Evaluation, and Recomendation



HANDLING IMBALANCED DATA USING SMOTE

	Accuracy		Precision		Recall		F1 Score	
	Before	After	Before	After	Before	After	Before	After
Logistic Regression	80.97%	78.42%	67.84%	59.02%	56.97%	67.25%	61.93%	62.87%
Xgboost	80.36%	79.51%	67.87%	62.09%	52.61%	63.07%	59.27%	62.58%
KNN	76.86%	70.71%	59.64%	47.25%	45.82%	67.42%	51.82%	55.56%
Random Forest	79.37%	79.08%	66.02%	63.12%	47.39%	52.79%	55.17%	57.5%
Gradien Boosting	80..64%	80.03%	68.46%	63.57%	53.31%	62.02%	59.90%	62.79%

MODELLING RESULT

Prediction and Evaluation

	Accuracy	Precision	Recall	F1 Score	ROC
Logistic Regression	80.12%	63.51%	63.07%	63.29%	74.76%
Gradien Boosting	79.98%	62.08%	67.6%	64.72%	76.98%
Xgboost	79.7%	62.56%	62.89%	62.73%	74.43%
Random Forest	79.56%	61.75%	64.98%	63.33%	74.99%
KNN	71.89%	48.73%	66.72%	56.32%	70.27%

RECOMENDATION

Model yang direkomendasikan yaitu Gradient Boosting yang memiliki:

- Accuracy : 79,98%
- Precision : 62.08%
- Recall : 67.7%
- F1 Score : 64.72%
- ROC : 76.98%
- Execution Time : 23 s

Conclusion

1. Gradient Boosting is the best model according to the performance of the model
2. Monthly charges are significantly affect the customer churn

Recommendation

1. Use feature selection, so it could improve the performance of the model
2. Retargeting to younger customer by creating the promo discount for longer term and other services
3. Increase the performance for all services to reduce customer churn and increase customer satisfaction

End

Thank you

DigitalSkola

Google Collab Link

EDA Link:

https://colab.research.google.com/drive/1V9Si_7DRGv7L5

DRta6AdnzikbRNWGBmk?usp=sharing

Modelling Link:

https://colab.research.google.com/drive/1F2TlXPT91-

6ITg2iHmAx4Gp8R4zj_EP8?usp=sharing