Customer Churn Analytics

Data Scientist Take Home Test - Telkom Indonesia

Customer Churn Data

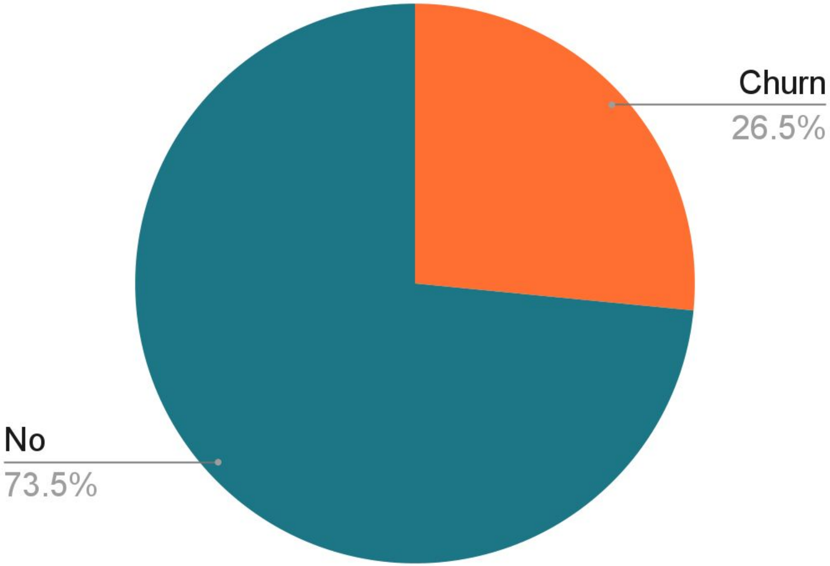
Features

customerID, gender, SeniorCitizen, Partner, Dependents, tenure, PhoneService, MultipleLines, InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies, Contract, PaperlessBilling, PaymentMethod, MonthlyCharges, TotalCharges,

Target Label

Churn

Customer Churn Statistics



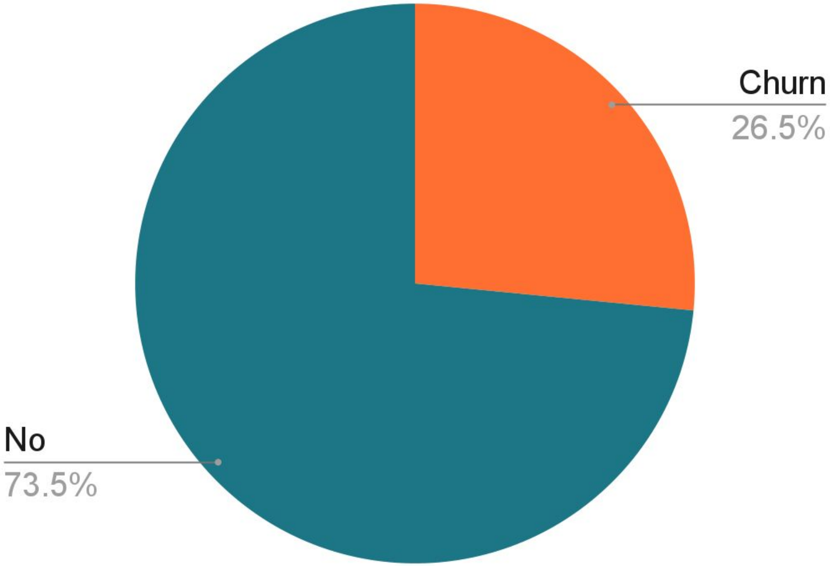
73.5%

No Churn

26.5%

Churn

Imbalance Data Problem



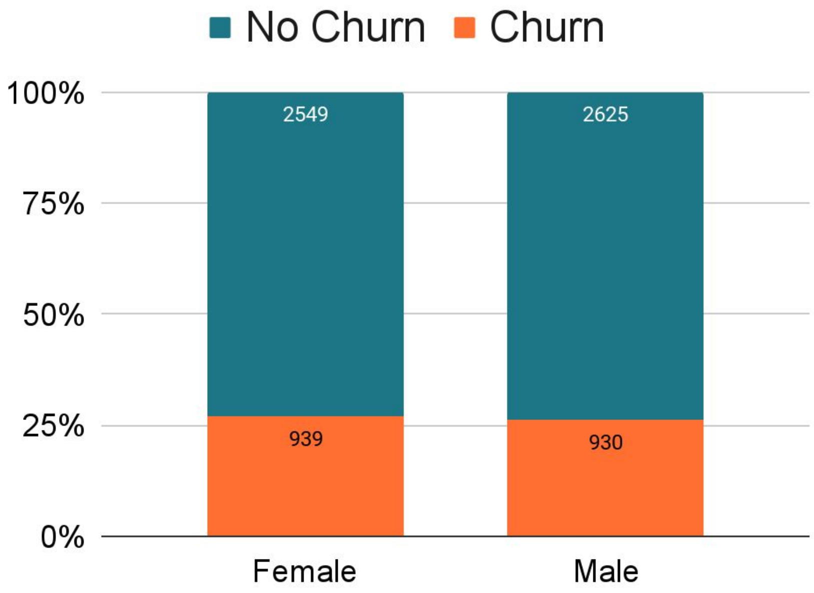
The target class has an **uneven distribution** of observations:

- Churn class label has a very low number of observations
- No Churn class label has a very high number of observations.

Analyzing Churn Likelihood

Analyzing Churn Likelihood

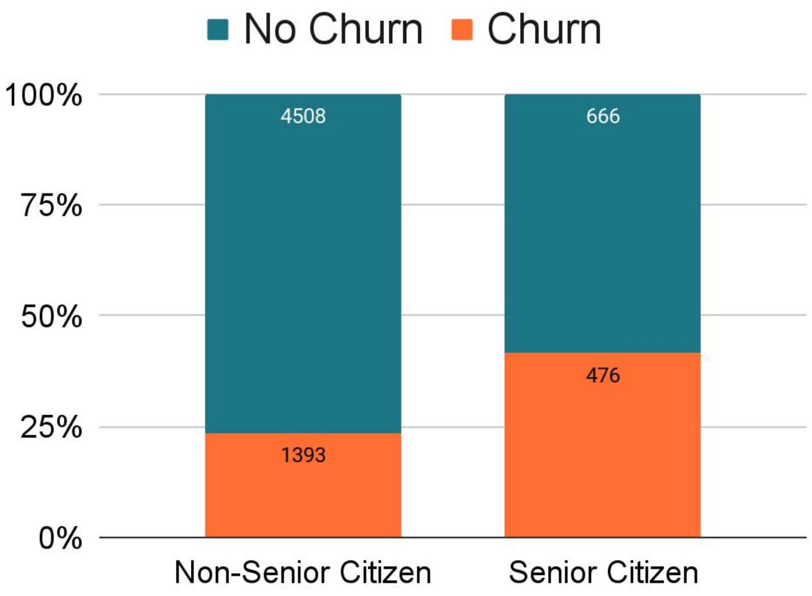
Churn Likelihood - Gender



Both male and female have a **similar proportion** of “Churn vs No Churn”, i.e. around 25% of them churn.

Thus, customer gender does not significantly affect the prediction whether a customer will churn or not.

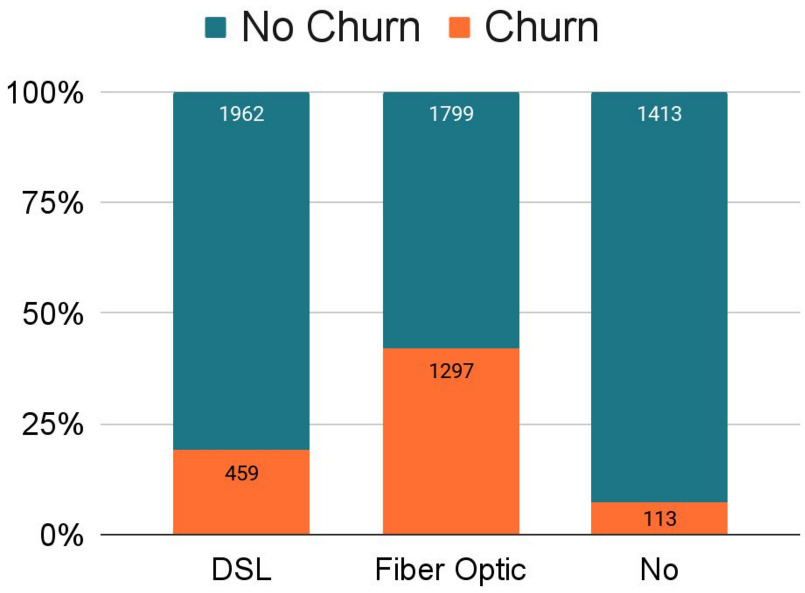
Churn Likelihood - Seniority of Citizen



Almost half of Senior Citizen churn, while only a quarter of Non-Senior Citizen churn.

Thus, Senior Citizen is more likely to churn compared to Non-Senior Citizen

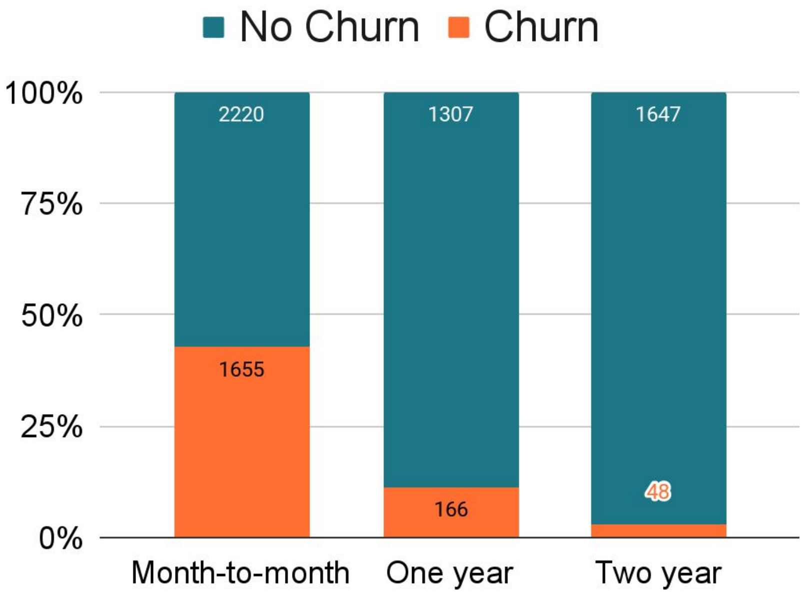
Churn Likelihood - Internet Service



Less than 25% of customers who use DSL and No internet service churn, while almost half of Fiber Optic customer churn.

Thus, Fiber Optic customers are more likely to churn compared to DSL and No Internet Service customers

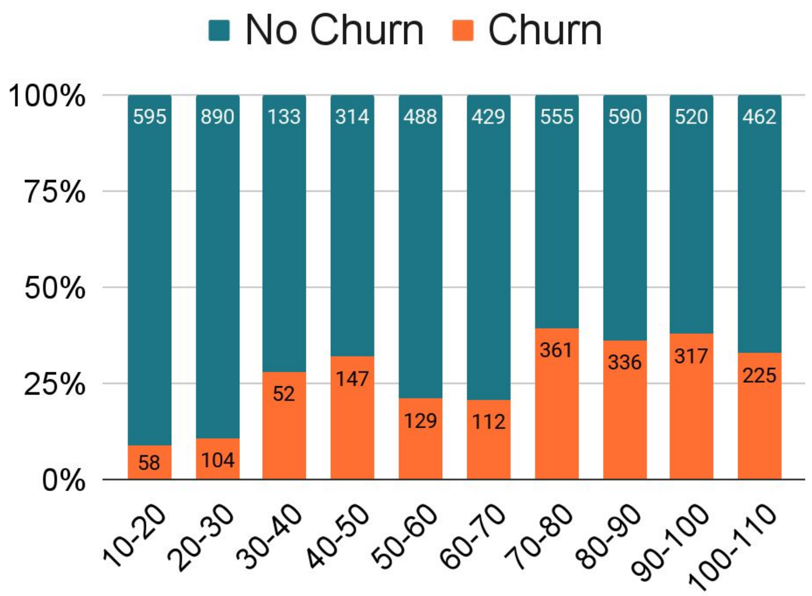
Churn Likelihood - Contract



The customers with month-to-month contract have the highest churn proportion. The customers with two year contract have the lowest churn proportion.

The longer customer contract will make the customer less likely to churn

Churn Likelihood - Monthly Charges



The monthly charges are binned into 10 bins with interval size 10. There is an uptrend pattern on the churn proportion when the monthly charges increases.

The customers who have a higher monthly charges are more likely to churn

Predicting using Machine Learning

Predicting using Machine Learning

Converting Categorical Feature into Numerical Feature

Categorical data are variables that contain label values rather than numeric values.
For example, there are 3 different internet service categories, i.e. fiber optic, DSL, and no service

ML algorithms require all input variables and output variables to be numeric.

Ordinal Encoding

Citizen Seniority
Senior Citizen
Non-Senior Citizen
Foreigner
Senior Citizen

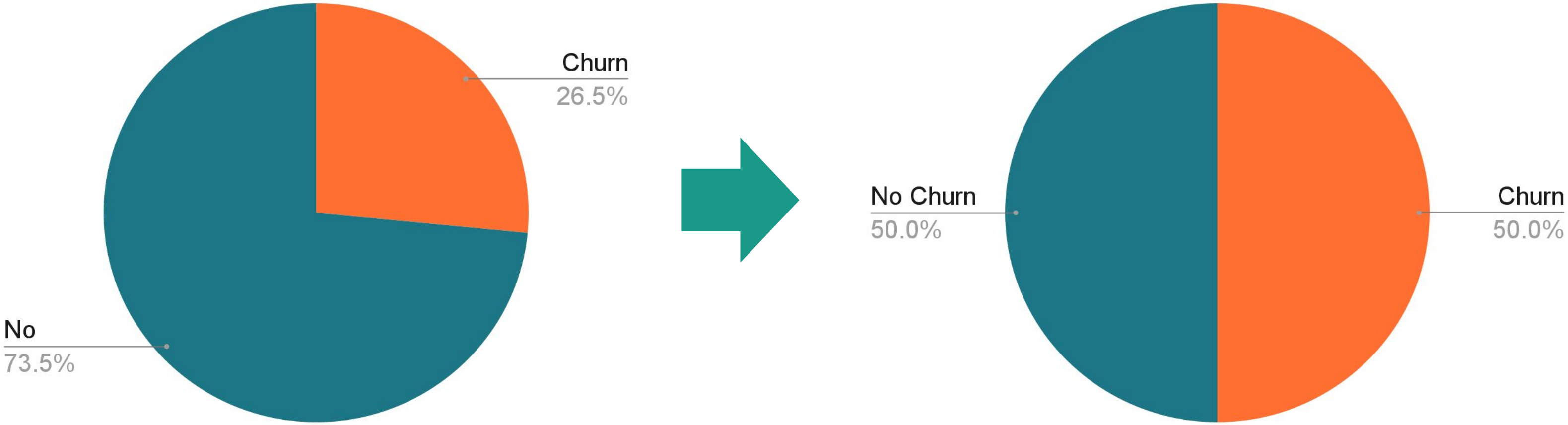
Citizen
1
0
2
1

One Hot Encoding

Inet Service
Fiber Optic
DSL
No
DSL

IS_FO	IS_DSL	IS_No
1	0	0
0	1	0
0	0	1
0	1	0

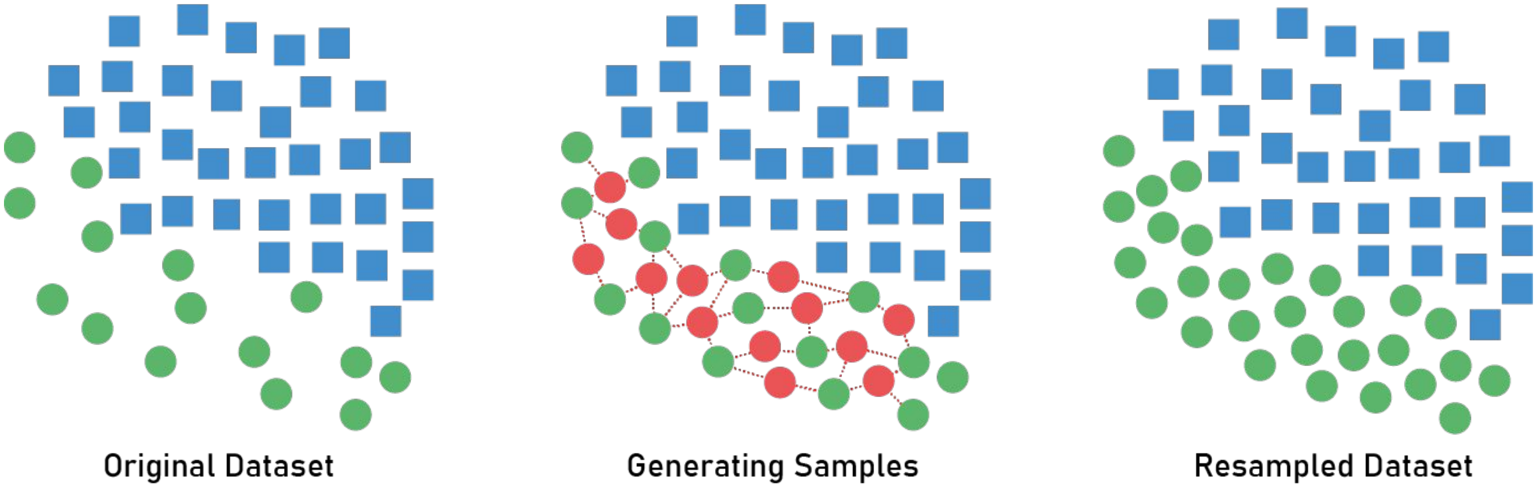
Handling Imbalance Data? Oversampling



Generate examples from the minority class, i.e. creating more churn data

SMOTE: Synthetic Minority Oversampling Technique

It generates new instances with the help of interpolation between the positive instances that lie together.



Hence, SMOTE helps to overcome the overfitting problem posed by random oversampling.

Appropriate Evaluation Metric for Imbalance Data

Accuracy Paradox

When one class of labels is underrepresented, it might get ignored while still keeping high accuracy of the model

F1 score

F1 score is very useful when dealing with imbalanced classes problems

$$F1\ score = 2 * \frac{Precision\ * Recall}{Precision + Recall}$$

Evaluating Machine Learning Models

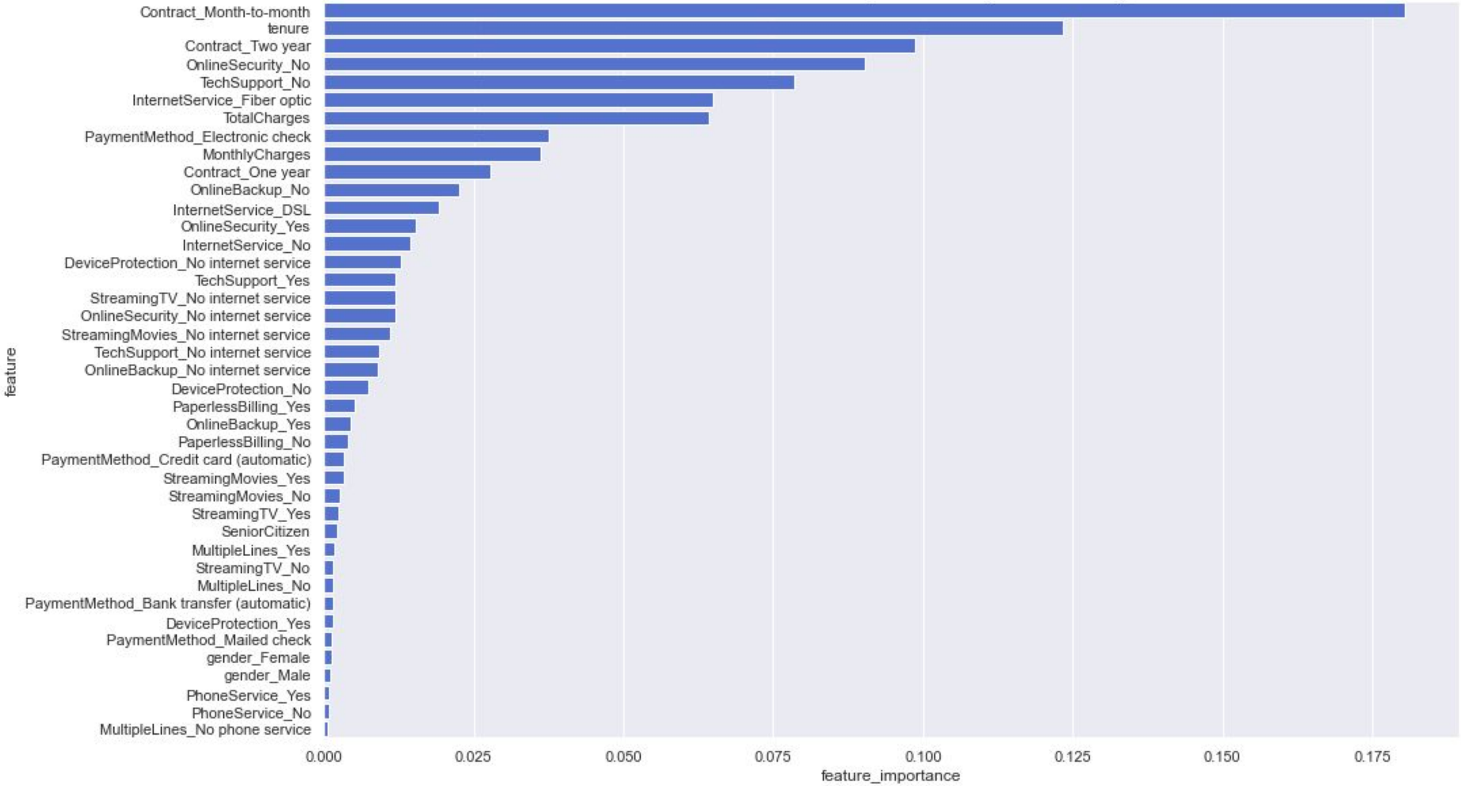
ML Model	Acc Train	Acc Test	F1 Train	F1 Test
Gaussian	0.77	0.77	0.78	0.78
kNN	0.83	0.76	0.83	0.78
MLP	0.75	0.75	0.73	0.73
RandomForest	**0.85**	**0.84**	**0.85**	**0.85**

RandomForest model achieves **the best F1 score** on both train and test data

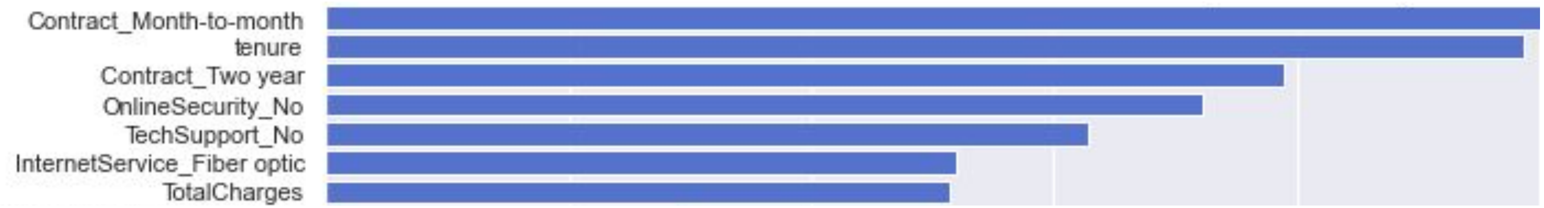
Insights and Recommendations

Insights and Recommendations

Analyzing Importance Features



Top 6 Importance Features



01 | Contract_Month-to-month

02 | tenure

03 | Contract_Two year

04 | OnlineSecurity_No

05 | TechSupport_No

06 | InternetService_Fiber Optic

Business Strategy

01 | Contract_Month-to-month

03 | Contract_Two year

>> | Upsell customers who subscribe month-to-month contract into a longer contract, e.g. two year contract

02 | tenure

>> | Add more discount in the first year or revise pricing strategy with marketing budget.
The goal is to prolong the user until they belong to a tenure bin where the churn rate is low (check next slide)

04 | OnlineSecurity_No

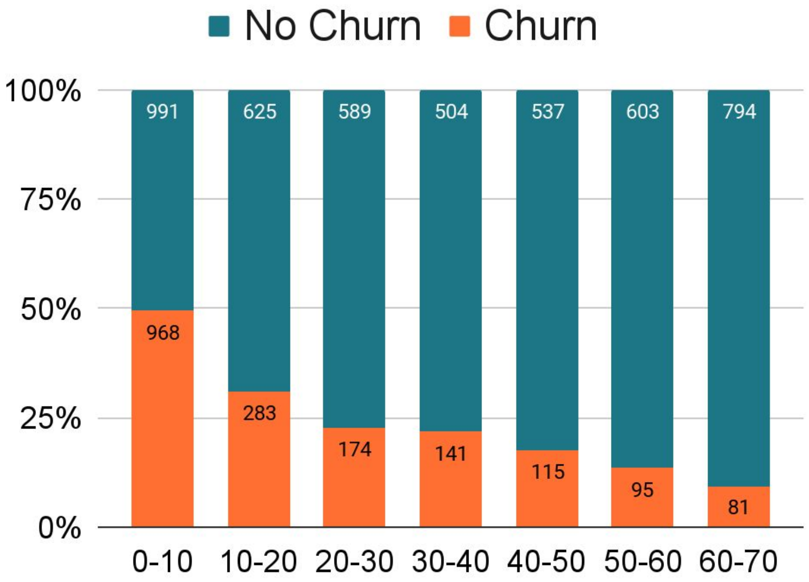
05 | TechSupport_No

>> | Offer bundled/discounted “tech support and online security” for those potentially upselling customers

06 | InternetService_Fiber Optic

>> | Consider widening the fibre optic outreach.
How to prioritize the outreach area? Reverse code from the customerID

Churn Likelihood - Tenure



The tenure are binned into 7 bins with interval size 10. There is a downtrend pattern on the churn proportion when the tenure increases.

The customer who has a higher tenure is less likely to churn.

About me!

My Personal Website: mhilmiasyrofi.github.io

Source Code: github.com/mhilmiasyrofi/CustomerChurnAnalytics