Purwadhika
Startup and Coding School

Telco Churn Predictive Analysis

CAPSTONE 3 PROJECT

Handika Eki Winata
JCDS 0306 Batam

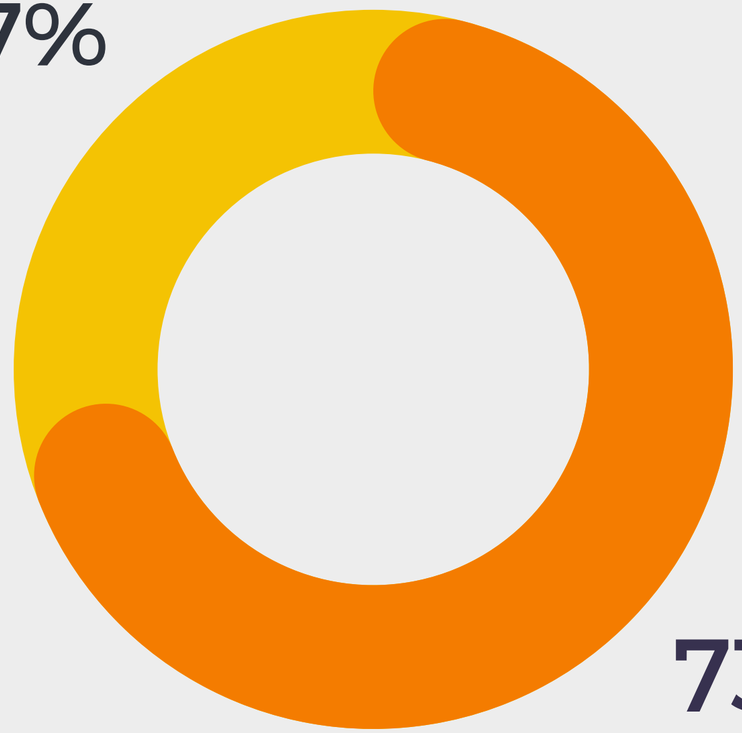
Table of Content

Background

- What is Churn? Why is it important in Telecomunication Industry?
- What are the factors driving customer to churn?
- The goal is to outline the patterns and trends that influence a churn, then construct a predictive model that can reliably detect prospective churn customer

Problem Statement

Performing a Customer Retention Cost Strategy with aims to Reduce Retention Cost Spent by lowering the Churn Percentages



"The cost of retaining an existing customer is far less than the cost of acquiring a new one"

Predictive Analysis Case

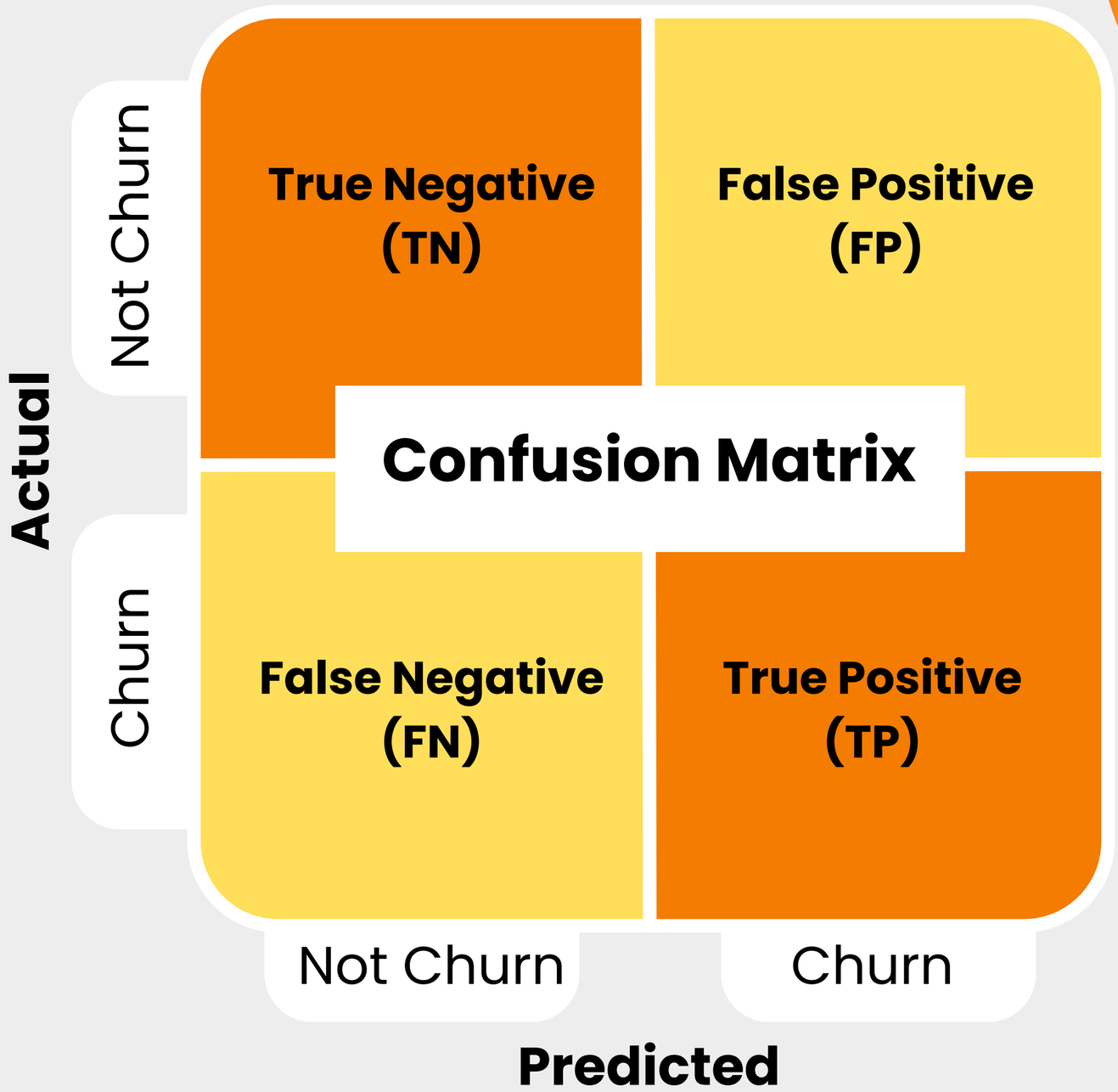
TARGET:

- 1 (Positive) = Churn / Turnover
- 0 (Negative) = Not Churn / Retained

FN > FP, since acquiring new customer is 5 times costly than retaining existing customers

Thus..

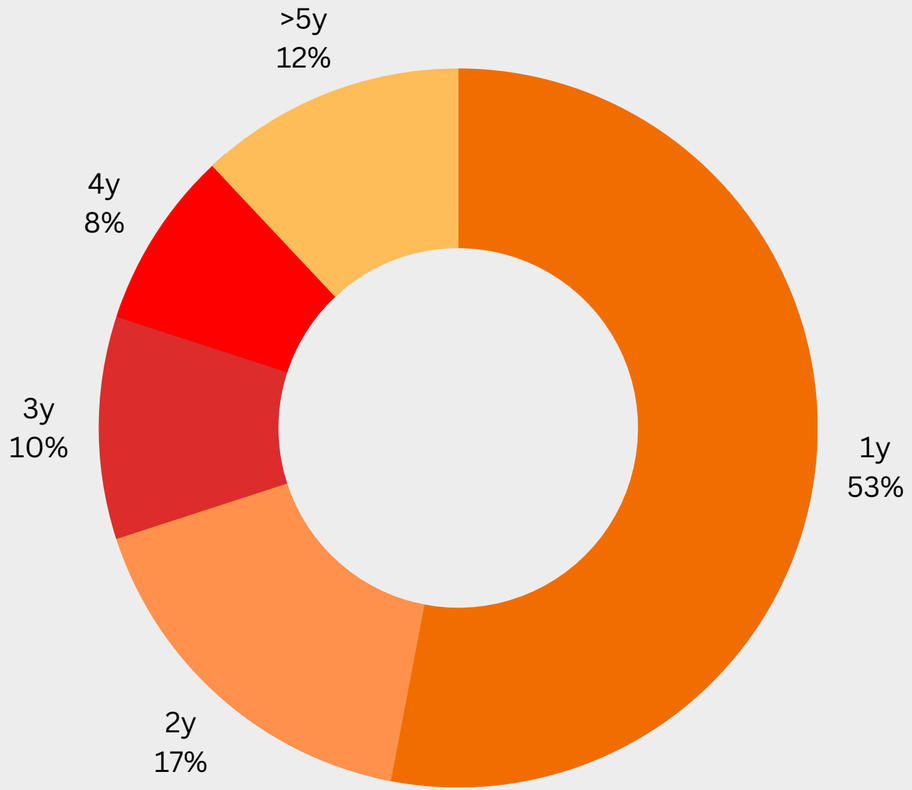
- **Recall** will be used as a baseline metrix to evaluate the models
- **Customer Retention Cost Strategy** will be used as a judge to chose which model is the best



EDA

#	Feature	Description
1	Dependents	Whether the customer has dependents or not
2	tenure	Number of months the customer has stayed with the company
3	OnlineSecurity	Whether the customer has online security or not
4	OnlineBackup	Whether the customer has online backup or not
5	InternetService	Whether the client is subscribed to Internet service
6	DeviceProtection	Whether the client has device protection or not
7	TechSupport	Whether the client has tech support or not
8	Contract	Type of contract according to duration
9	PaperlessBilling	Bills issued in paperless form
10	MonthlyCharges	Amount of charge for service on monthly bases
11	Churn	1 if the customer Churn, 0 otherwise

Churn Customer based on Tenure Category



- Most of the Churn customer = tenure of below 1 year
- Considered as new customer
- Possibly due to unsatisfied in our service or better offers by the competitors

ML Modeling

PREPROCESSING

Data Splitting

Train Test Ratio = 75% : 25%
Stratify = y

Scaling

MinMaxScaler

Encoding

OneHotEncoding
Ordinal Encoding

Model Benchmarking

Logistic Regression
K-Nearest Neighbor
Decision Tree
Random Forest
Gradient Boosting
XGBoost

MODEL

Balanced Accuracy
Recall
Precision
F1 Score

METRICS

RESAMPLING

None
Random Under Sampling
Random Over Sampling
Near Miss
SMOTE
SMOTEENN

CROSS VALIDATE

Stratified K-Fold

	model	resample	accuracy	precision	recall	F1
31	Logistic Regression	smoteenn	75.934546	50.065667	81.161263	61.903655
17	Gradient Boosting	rus	75.858471	50.179740	80.747289	61.867449
13	Logistic Regression	rus	75.712100	50.231395	80.229689	61.745296
7	Logistic Regression	ros	76.078991	51.030673	79.916671	62.262013
19	Logistic Regression	smote	75.667846	51.338405	78.158752	61.942168

- Model benchmarking in respect to the `Recall` metric
- `Logistic Regression` is dominating the top 5 of model ranking.
- Gradient Boosting achieve the 2nd place.
- The top 3 model will be tested again together with 2 resampling method `SMOTEENN` and `RandomUnderSampling` to predict the test data
- The prediction result will be used in determining the best model

Benchmarking Result

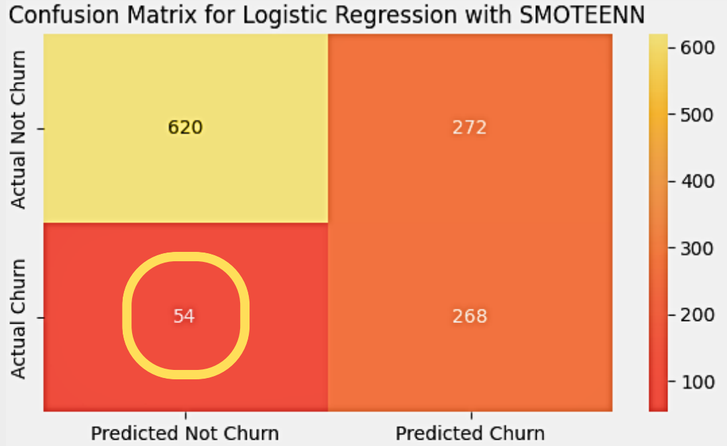
Prediction Result

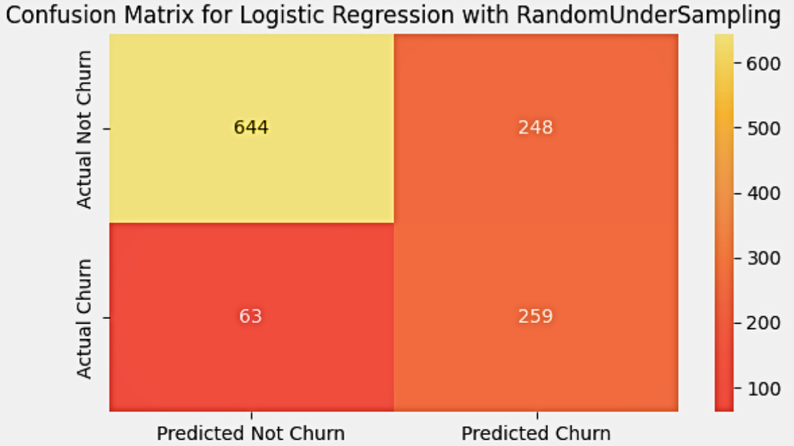
	Accuracy	Precision	Recall	F1
Logistic Regression with SMOTEENN	73.146623	49.629630	83.229814	62.180974
Logistic Regression with RandomUnderSampling	74.382208	51.084813	80.434783	62.484922
Gradient Boosting with SMOTEENN	74.958814	51.890756	76.708075	61.904762
Gradient Boosting with RandomUnderSampling	73.805601	50.392157	79.813665	61.778846

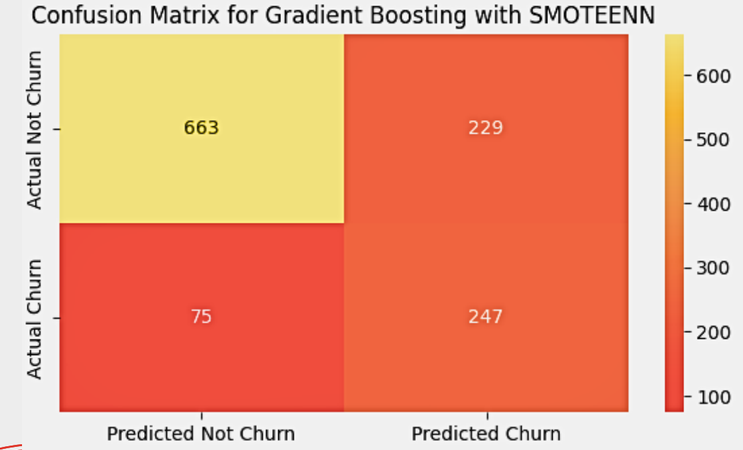
Best Model: Logistic Regression

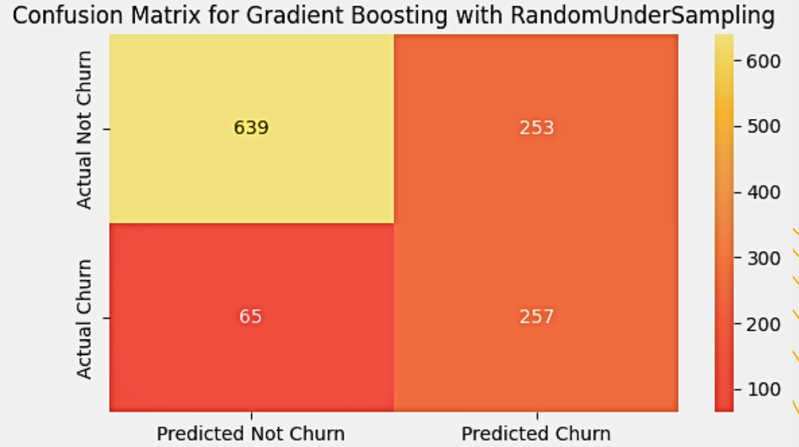
- Resampling = SMOTEENN
- Recall = 83%
- FN = 54

Let's further increase the `Recall` percentage by performing the hyperparameter tuning to the chosen `Best Model`*









1st Tuning

Hyperparam space:

penalty = [None, l1, l2, 'elastic-net]

C = [0.001, 0.01, 0.1, 1, 19, 100]

max_iter = [100, 500, 1000]

solver = [liblinear, saga, lbfgs]

Best Param:

Penalty = l1, C = 0.01,

Max iter = 100Solver = saga

Best Score: 0.90%

2nd Tuning

Hyperparam space:

penalty = [None, l1, l2, 'elastic-net]

C = [0.001, 0.0.005, 0.01, 0.05, 0.1]

max_iter = [50, 75, 100, 125, 150]

solver = [liblinear, saga, lbfgs]

Best Param:

Penalty = l1, C = 0.005,

Max iter = 150, Solver = saga

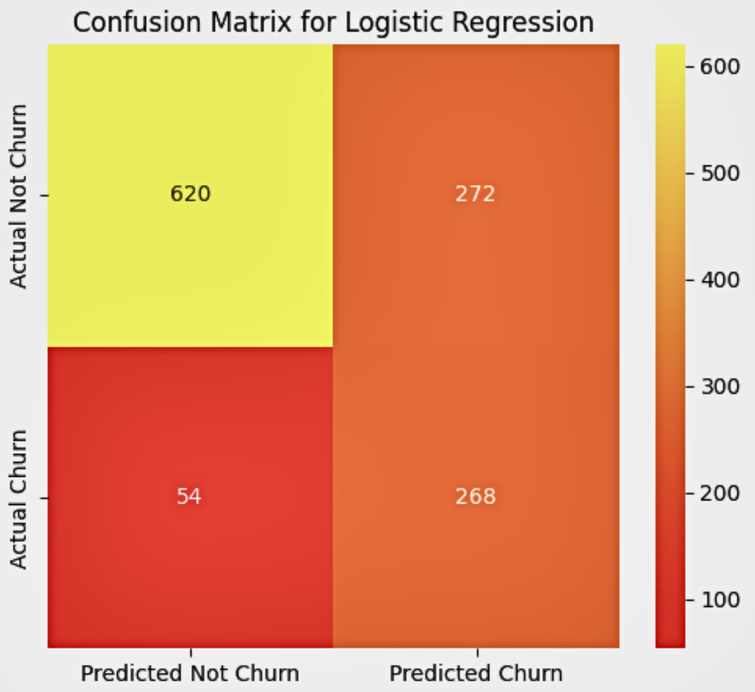
Best Score: 0.94%

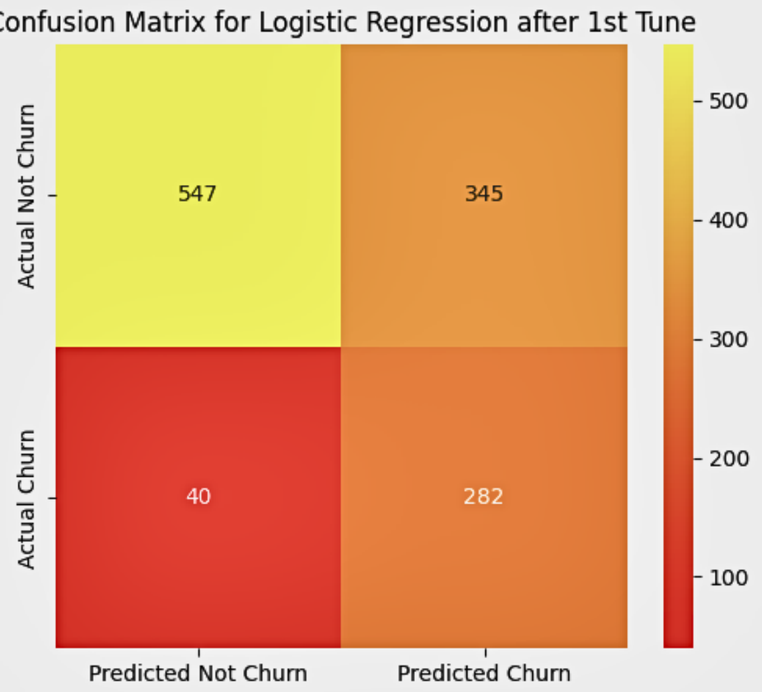
Hyperparameter Tuning

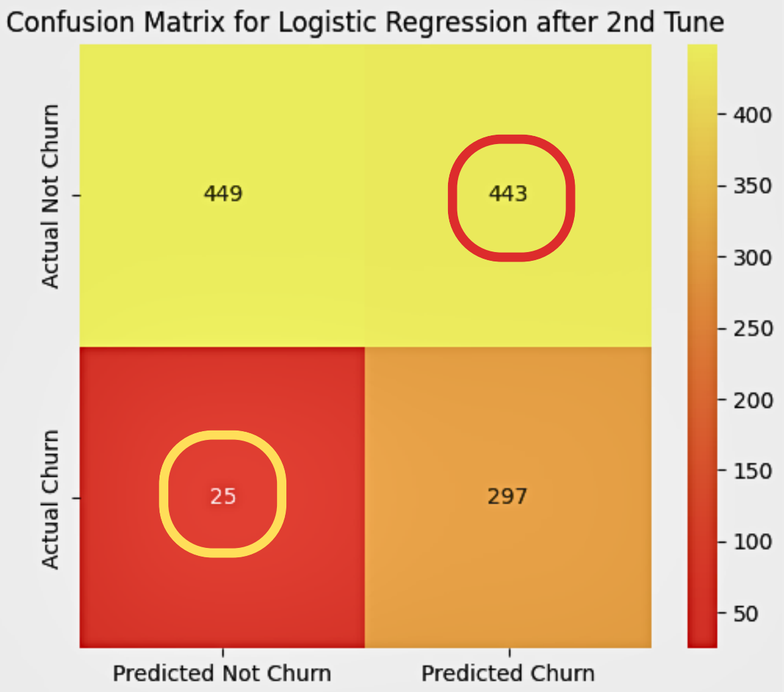
With Grid Search - Cross Validate

	Accuracy	Precision	Recall	F1
Logistic Regression 2nd Tune	61.449753	40.135135	92.236025	55.932203
Logistic Regression After 1st Tune	68.286656	44.976077	87.577640	59.430980
Logistic Regression before Tune	73.146623	49.629630	83.229814	62.180974

- Recall is successfully increased from 83% to 92%
- With FN number drop drastically
- But Sacrificing FP to rise too high







Tuning Result

In Predicting Test data

Case Study

Customer Retention Cost (CRC) The cost of acquiring new customers is five times higher than the cost of retaining existing customers

Concept

Retention Cost = $1 -> TP, FP

Acquisition Cost = $5 -> FN

Formula

CRC Without a model:

- $1 x (TP + FP + TN + FN)

CRC With a ML Model:

- (FN x $5) + (TP x $1) + (FP x $1)

Calculation

CRC Results:

- Without a model = $1,214
- Logistic Regression before Tune = $810
- Logistic Regression after 1st Tune = $827
- Logistic Regression after 2nd Tune = $865

Percentage of CRC reduced by the model

$$Cost\ Spent = \left(\frac{Without\ model\ -\ With\ chosen\ model}{Without\ model} \right)$$

$$Cost\ Spent = \left(\frac{\$1,214\ -\ \$810}{\$1,214} \right) \approx 33\%$$

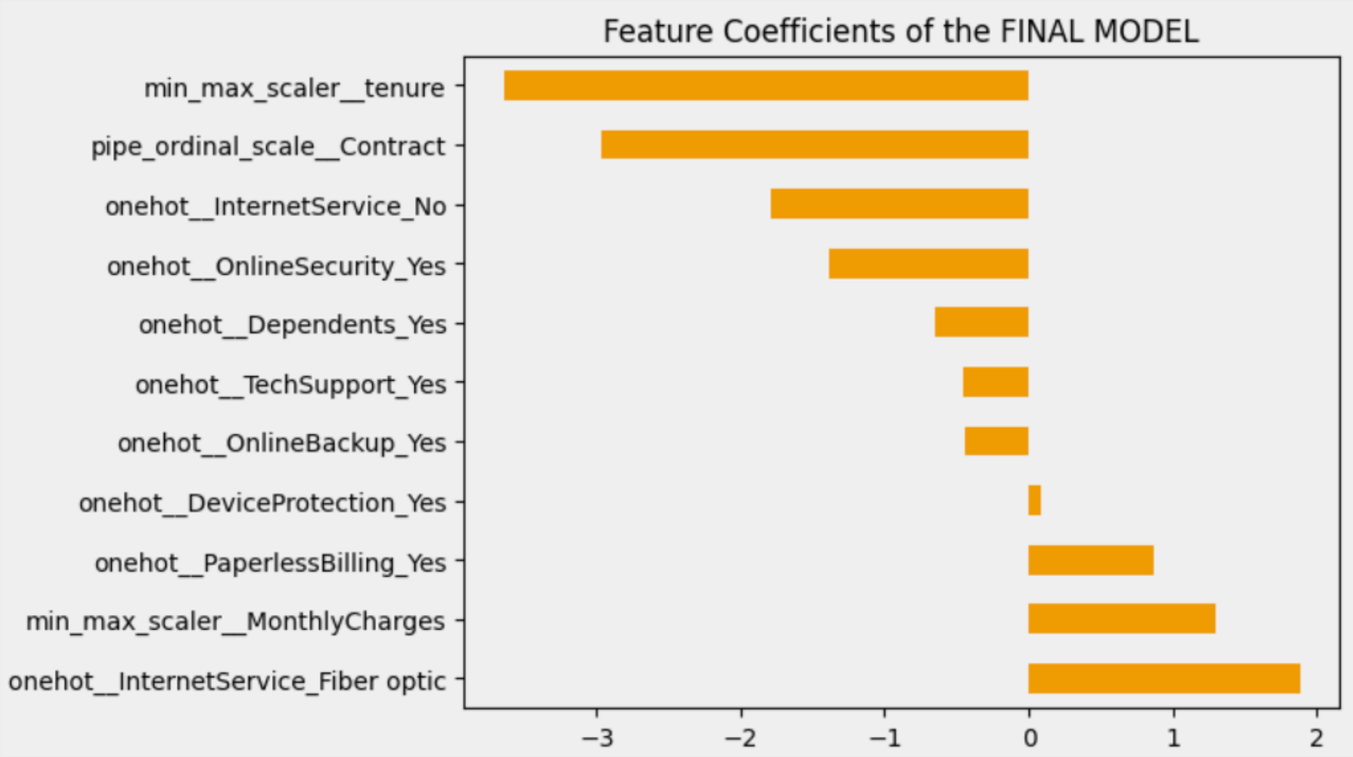
Verdict

Final Model:

- Logistic Regression before Tune

Feature Importance

Customer Retention Cost (CRC)



Negative Influence:

- Tenure, Contract - The longer tenure & contract decrease a chance of customer to Churn
- Internet service - No
- Having Dependent circumstance, Tech Support, Online Security & Backup - Reflect satisfaction

Positive Influence:

- Fiber Optic - Reflect dissatisfaction
- Monthly Charges - Higher Charges increase a chance of customer to Churn
- Paperless Billing, Device Protection - Yes

RECOMMENDATION

Data

- Customer Demographics
- Churn Rate Calculation
- Seasonal Usage Pattern
- Competitor Analysis

ML Model

- Continuous Monitoring and Updating
- Experiment with different algorithm
- Deployment in Real-Time System
- Model Interpretability tools

Business

- Customer Retention Strategy, Loyalty Program
- Rewarding the Returned Customer
- Improve Fiber Optic and Network Quality service
- Proactive Customer Service
- Feedback and Reviews

CONCLUSION

Final Model: *"Logistic Regression before Tune"*

- Most Cost-Effective Model
- Reduce Retention Cost by 33%
- Recall (87%) and FN (54/322)

Service to improve:

- Fiber Optic and Network Quality
- Rate of Monthly Charges
- Paperless Billing and Device Protection

Thank You

CAPSTONE 3 PROJECT
Handika Eki Winata
JCDS 0306 Batam