BINAR

Telecom Churn Prediction

Table of Content

BINAR

Problem

Customer churn merupakan salah satu permasalahan yang mempengaruhi keberlanjutan bisnis dalam industri telekomunikasi. Biaya mengakuisisi pelanggan baru jauh lebih tinggi daripada mempertahankan pelanggan yang sudah ada. Maka dari itu, mengidentifikasi fitur yang berpengaruh signifikan terhadap *customer churn* sekaligus memprediksi *customer churn* penting dilakukan agar dapat memberikan aksi yang tepat sesuai kebutuhan pelanggan sehingga dapat mencegah pelanggan untuk *churn*.

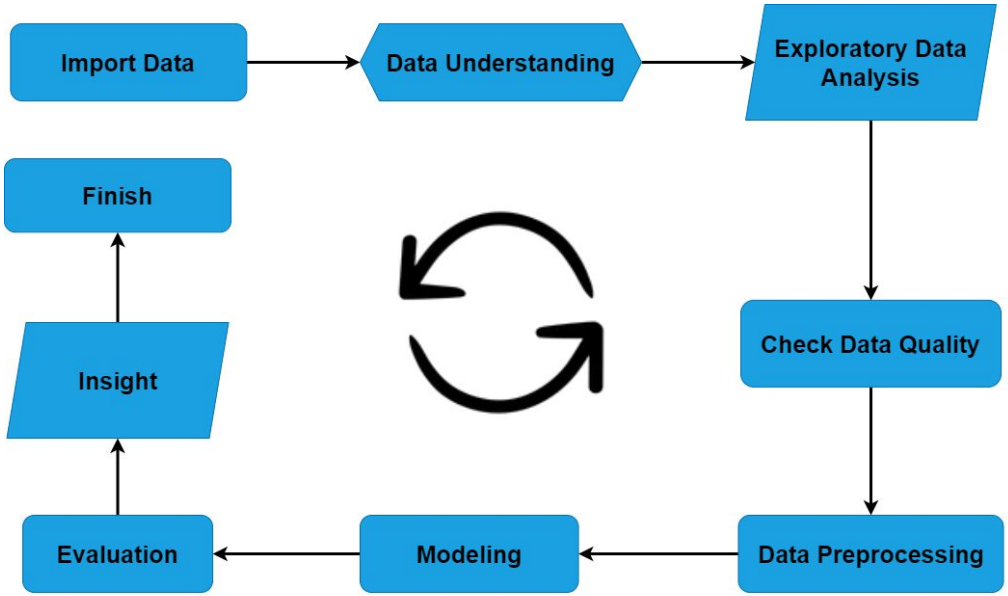
Goal

01 Mengetahui penyebab umum pelanggan berhenti menggunakan layanan telekomunikasi.

02 Mencari model klasifikasi terbaik untuk memprediksi status *churn* pelanggan layanan telekomunikasi.

03 Memprediksi status *churn* pelanggan layanan telekomunikasi pada *unseen* data.

Analysis Stage



Drop variabel state

Encoding

Splitting

Standarization

Imbalance Handling

Data Understanding

Pada penelitian ini menggunakan data sekunder yang diperoleh dari *Kaggle*, yaitu *Telecom Churn Dataset*. *Telecom Churn Dataset* adalah dataset yang digunakan untuk memprediksi pelanggan yang akan berhenti menggunakan layanan telekomunikasi. Dataset ini berfokus pada masalah *churn*, yaitu pelanggan yang mengakhiri langganan layanan telekomunikasi. Selain itu, dataset ini terdiri dari data training dan data testing, dimana pelatihan model akan dilakukan pada data train. Selanjutnya, model terbaik yang telah diperoleh pada tahap pelatihan digunakan untuk memprediksi status *churn* pelanggan pada data test.

Data Understanding

Variabel	Tipe Data	Keterangan
State	Nominal	kolom yang mencantumkan tempat pelanggan berada.
account_length	Numerik	lamanya waktu pelanggan memiliki akun dengan perusahaan telekomunikasi.
area_code	Nominal	kode area yang terkait dengan nomor telepon pelanggan.
international_plan	Nominal	menunjukkan apakah pelanggan memiliki paket panggilan internasional.
voice_mail_plan	Nominal	menunjukkan apakah pelanggan memiliki paket *voicemail*
number_vmail_messages	Numerik	Berisi jumlah pesan *voicemail* yang dimiliki pelanggan.
Total_day_minutes Total_day_calls total_day_charge	Numerik	Ketiga kolom ini mencatat informasi tentang panggilan siang hari, termasuk total menit, jumlah panggilan, dan total biaya.

Data Understanding

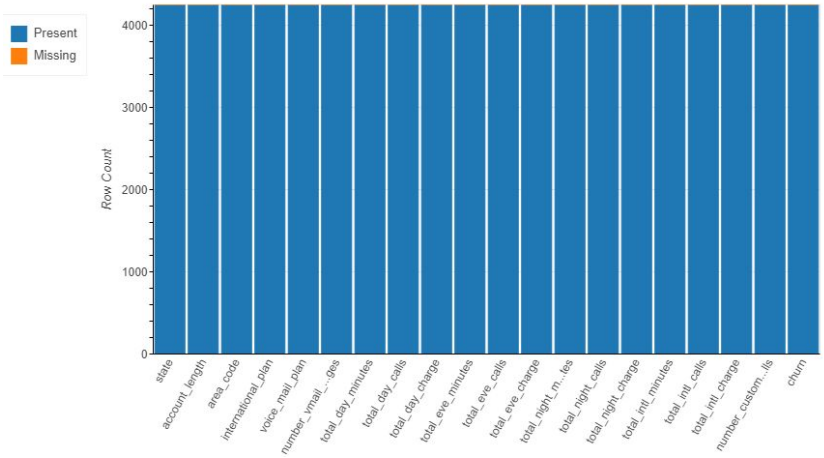
BINAR

Variabel	Tipe Data	Keterangan
Total_eve_minutes Total_eve_calls total_eve_charge	Numerik	Ketiga kolom ini mencatat informasi tentang panggilan petang hingga tengah malam, termasuk total menit, jumlah panggilan, dan total biaya.
Total_night_minutes Total_night_calls total_night_charge	Numerik	Ketiga kolom ini mencatat informasi tentang panggilan malam hari hingga sebelum dini hari, termasuk total menit, jumlah panggilan, dan total biaya.
Total_intl_minutes Total_intl_calls total_intl_charge	Numerik	Ketiga kolom ini mencatat informasi tentang panggilan internasional, termasuk total menit, jumlah panggilan, dan total biaya.
number_customer_service_calls	Numerik	mencatat jumlah panggilan layanan pelanggan yang dibuat oleh pelanggan.
churn	Nominal	Kolom ini mencatat apakah pelanggan telah meninggalkan layanan.

Exploratory Data Analysis

Check Data Quality

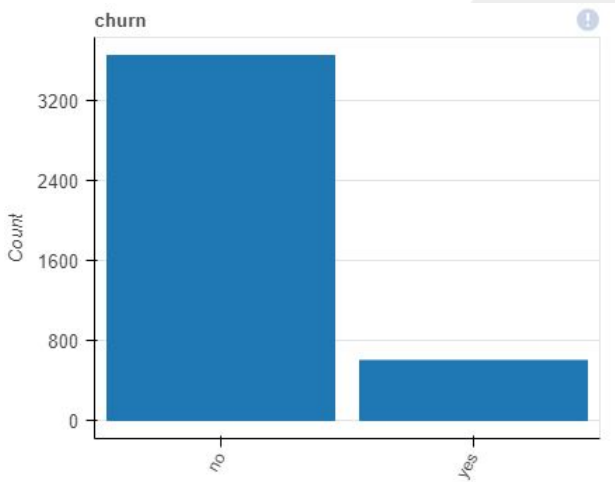
Plot Missing value



Duplicate Data

Duplicate Rows	0
Duplicate Rows (%)	0.0%

Barplot churn

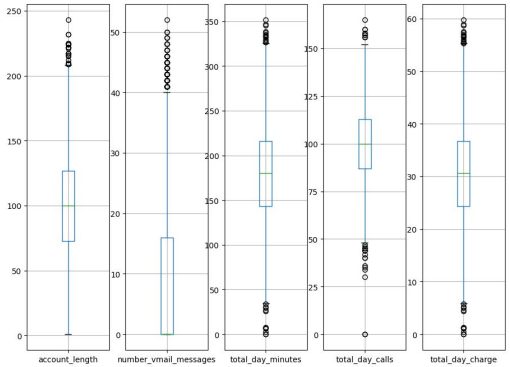


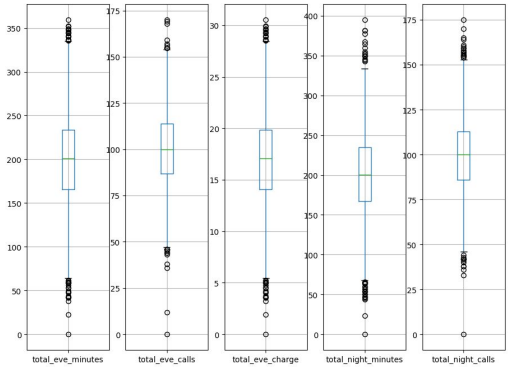
Dataset ini tidak memiliki *missing value* dan data duplikat. Namun, terdapat kelas yang tidak seimbang. Untuk analisis selanjutnya, penulis menggunakan Teknik Sampling SMOTEENN untuk mengatasi ketidakseimbangan kelas ini.

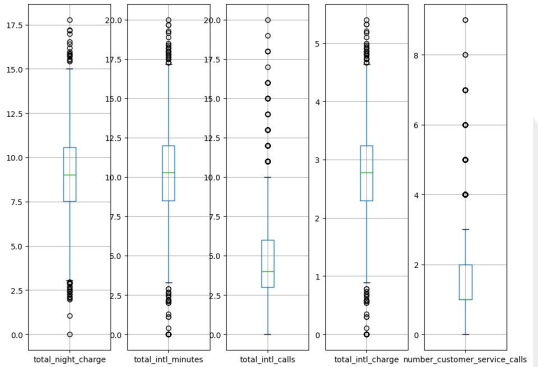
Exploratory Data Analysis

Check Data Quality

Boxplot Variabel Numerik







Jumlah Outlier pada Setiap Variabel Numerik

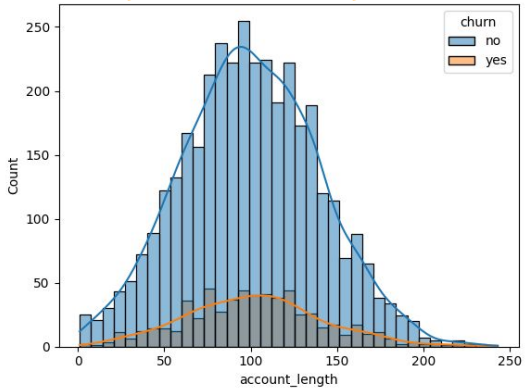
```
find_outliers_IQR(train[num]).notnull().sum()

account_length                20
number_vmail_messages         86
total_day_minutes             25
total_day_calls               28
total_day_charge              26
total_eve_minutes             34
total_eve_calls               24
total_eve_charge              34
total_night_minutes           37
total_night_calls             33
total_night_charge            37
total_intl_minutes            62
total_intl_calls             100
total_intl_charge             62
number_customer_service_calls 335
dtype: int64
```

Variabel numerik cenderung memiliki outlier. Misalkan setiap variabel numerik kita jumlahkan jumlah outliernya, maka proporsi outlier pada jumlah seluruh data dalam dataset sekitar 22.2%. Maka dari itu, **agar tidak kehilangan banyak data dan meninjau bahwa dataset ini tidak terlalu memiliki banyak data, penulis mengatasi data outlier pada tingkat algoritma, dimana akan digunakan algoritma klasifikasi yang secara alami dapat mereduksi efek buruk outlier.** Algortima klasifikasi ini akan dibahas pada slide-slide berikutnya.

Exploratory Data Analysis

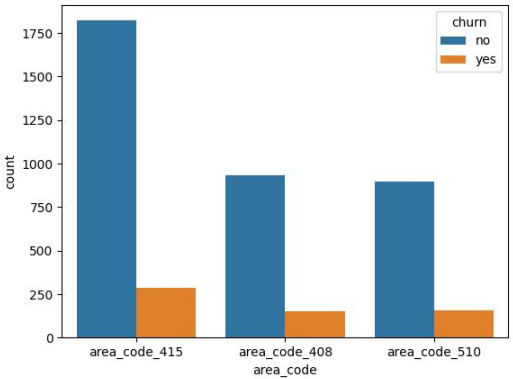
Insight based on Descriptive Statistics

Histogram *Account Length* vs *Churn*



Secara deskriptif, tidak terdapat perbedaan yang signifikan antara pelanggan dengan status *churn* dan tidak *churn* dalam hal lamanya waktu pelanggan memiliki akun dengan perusahaan telekomunikasi. Hal ini dapat dilihat dari distribusi variabel *account length*, mean, dan median yang cenderung sama baik untuk pelanggan dengan status *churn* dan tidak *churn*.

Barplot *Area Code* vs *Churn*

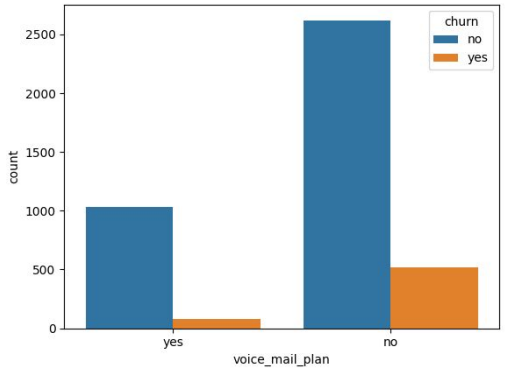


Pada dataset ini pelanggan terbanyak berada pada kode area 415. Namun, proporsi *churn* terbanyak berada pada kode area 510. Secara deskriptif, terlihat tidak terdapat perbedaan yang signifikan terhadap proporsi *churn* pada masing-masing kode area, dimana kode area 408, kode area 415, dan kode area 510 memiliki proporsi *churn* terhadap jumlah pelanggan pada area tersebut secara berurutan 14%, 13.6% dan 15%.

Exploratory Data Analysis

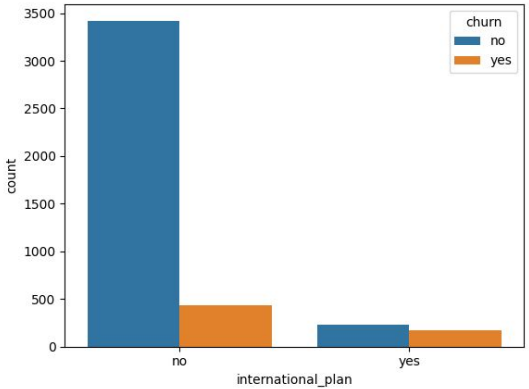
Insight based on Descriptive Statistics

Barplot *Voice Mail Plan* vs *Churn*



Kebanyakan pelanggan tidak memiliki paket *voice mail*. Secara deskriptif, pelanggan yang tidak memiliki paket *voice mail* cenderung berpotensi untuk *churn*. Dimana proporsi *churn* pada pelanggan yang tidak memiliki paket *voice mail* dan pelanggan yang memiliki paket *voice mail*, secara berurutan 16.4% dan 7.4%

Barplot *International Plan* vs *Churn*

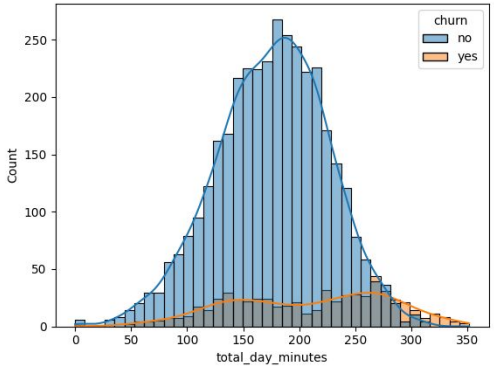


Kebanyakan pelanggan tidak memiliki paket panggilan internasional. Namun, pelanggan yang memiliki paket panggilan internasional memiliki tingkat *churn* yang lebih tinggi daripada pelanggan yang tidak memiliki paket panggilan internasional, dimana proporsi *churn* pada pelanggan yang memiliki paket panggilan internasional dan pelanggan yang tidak memiliki paket panggilan internasional secara berurutan, 42.2% dan 11.2%

Exploratory Data Analysis

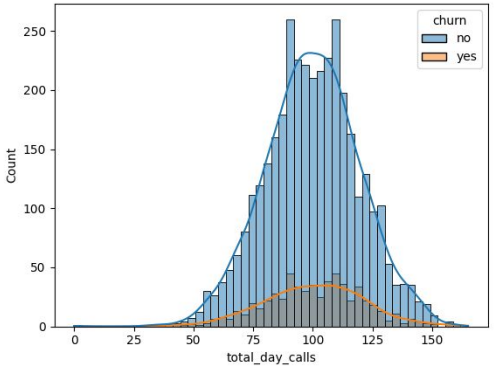
Insight based on Descriptive Statistics

Histogram *Total Day Minutes* vs *Churn*



Berdasarkan histogram disamping, pelanggan – pelanggan dengan jumlah panggilan siang hari dalam menit yang cenderung tinggi akan berpeluang untuk *churn*. Selanjutnya, ditinjau dari mean atau median pada variabel *total day minutes*, pelanggan yang *churn* cenderung memiliki mean dan median yang lebih tinggi dari pada pelanggan yang tidak *churn*.

Histogram *Total Day Calls* vs *Churn*

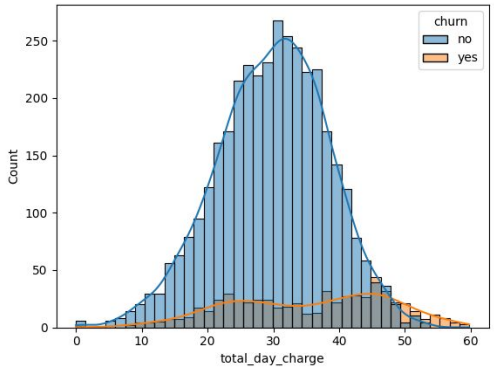


Secara deskriptif, tidak terdapat perbedaan yang signifikan antara pelanggan dengan status *churn* dan tidak *churn* dalam jumlah panggilan siang hari. Hal ini dapat dilihat dari distribusi variabel *total day calls*. Selain itu, mean dan median pada variabel ini cenderung memiliki nilai yang sama baik untuk pelanggan dengan status *churn* dan tidak *churn*.

Exploratory Data Analysis

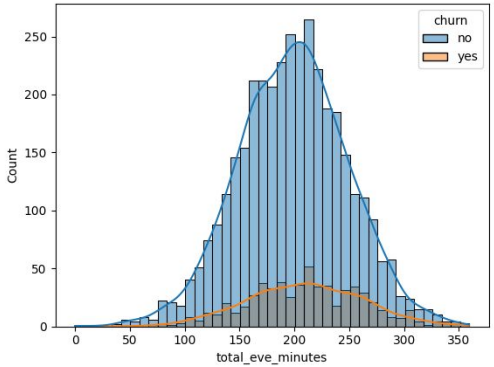
Insight based on Descriptive Statistics

Histogram *Total Day Charge* vs *Churn*



Berdasarkan histogram dibawah, pelanggan – pelanggan dengan total biaya panggilan siang hari yang cenderung mahal akan berpeluang untuk *churn*. Selanjutnya, ditinjau dari mean atau median pada variabel *total day charge*, pelanggan yang churn cenderung memiliki mean dan median yang lebih tinggi dari pada pelanggan yang tidak *churn*.

Histogram *Total Eve Minutes* vs *Churn*

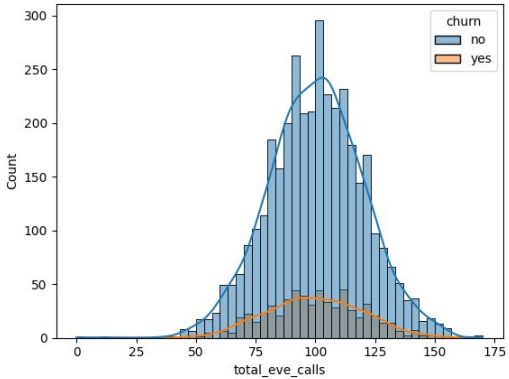


Secara deskriptif, tidak terdapat perbedaan yang signifikan antara pelanggan dengan status *churn* dan tidak *churn* dalam jumlah menit panggilan dari petang hingga tengah malam. Hal ini dapat dilihat dari distribusi variabel *total eve minutes*. Selain itu, mean dan median pada variabel ini cenderung tidak berbeda signifikan baik untuk pelanggan dengan status *churn* dan tidak *churn*.

Exploratory Data Analysis

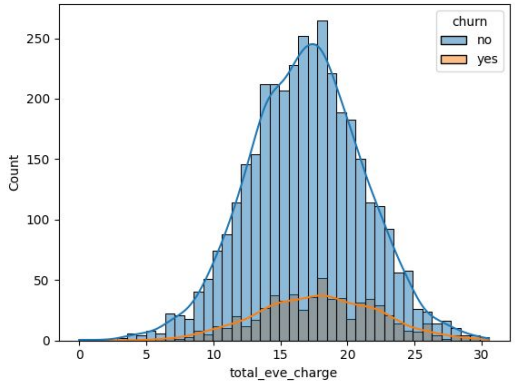
Insight based on Descriptive Statistics

Histogram *Total Eve Calls* vs *Churn*



Secara deskriptif, tidak terdapat perbedaan yang signifikan antara pelanggan dengan status *churn* dan tidak *churn* dalam hal banyak panggilan dari petang hingga tengah malam. Hal ini dapat dilihat dari distribusi variabel *total eve calls*, mean, dan median yang cenderung sama baik untuk pelanggan dengan status *churn* dan tidak *churn*.

Histogram *Total Eve Charge* vs *Churn*

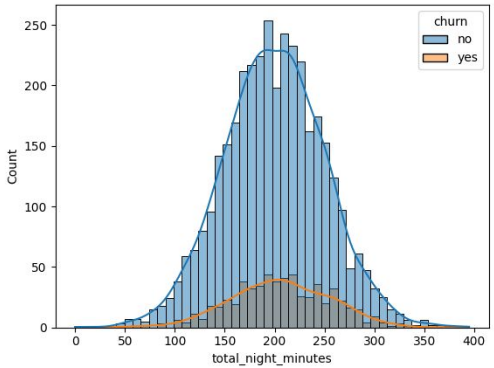


Secara deskriptif, tidak terdapat perbedaan yang signifikan antara pelanggan dengan status *churn* dan tidak *churn* dalam hal biaya panggilan yang dikeluarkan dari petang hingga tengah malam. Hal ini dapat dilihat dari distribusi variabel *total eve charge*, mean, dan median yang cenderung sama baik untuk pelanggan dengan status *churn* dan tidak *churn*.

Exploratory Data Analysis

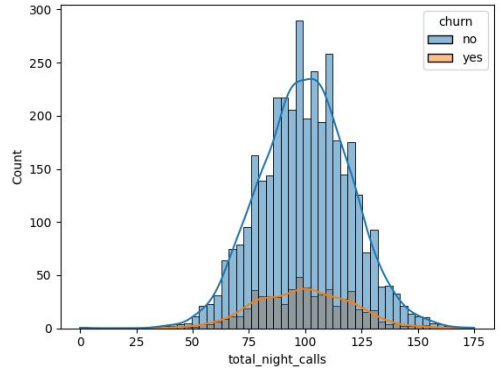
Insight based on Descriptive Statistics

Histogram *Total Night Minutes* vs *Churn*



Secara deskriptif, tidak terdapat perbedaan yang signifikan antara pelanggan dengan status *churn* dan tidak *churn* dalam hal jumlah panggilan dari tengah malam hingga sebelum dini hari dalam menit. Hal ini dapat dilihat dari distribusi variabel total night minutes yang cenderung sama untuk status *churn* dan tidak *churn*.

Histogram *Total Night Calls* vs *Churn*

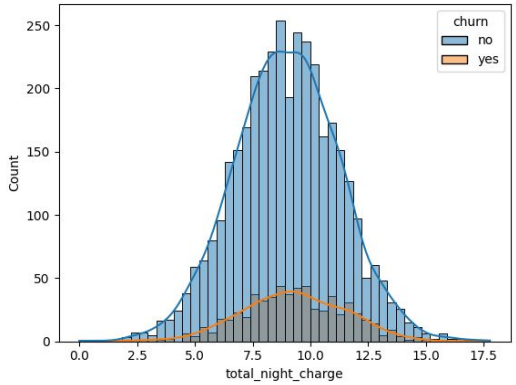


Secara deskriptif, tidak terdapat perbedaan yang signifikan antara pelanggan dengan status *churn* dan tidak *churn* dalam hal banyak panggilan dari tengah malam hingga sebelum dini hari. Hal ini dapat dilihat dari distribusi variabel *total night calls*, mean, dan median yang cenderung sama baik untuk pelanggan dengan status *churn* dan tidak *churn*.

Exploratory Data Analysis

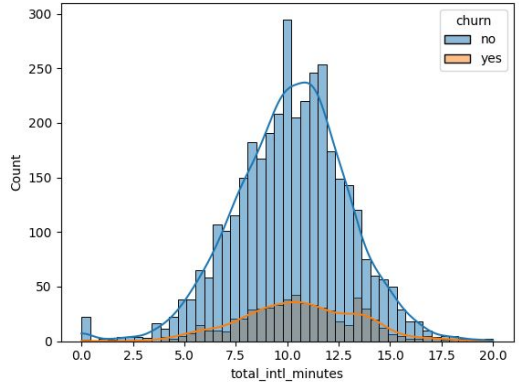
Insight based on Descriptive Statistics

Histogram *Total Night Charge* vs *Churn*



Secara deskriptif, tidak terdapat perbedaan yang signifikan antara pelanggan dengan status *churn* dan tidak *churn* dalam hal biaya panggilan dari tengah malam hingga sebelum dini hari yang dikeluarkan. Hal ini dapat dilihat dari distribusi variabel *total night charge*, mean, dan median yang cenderung sama baik untuk pelanggan dengan status *churn* dan tidak *churn*.

Histogram *Total intl Minutes* vs *Churn*

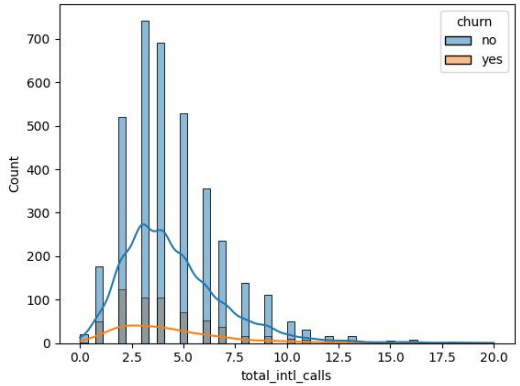


Secara deskriptif, tidak terdapat perbedaan yang signifikan antara pelanggan dengan status *churn* dan tidak *churn* dalam hal panggilan internasional dalam menit. Hal ini dapat dilihat dari distribusi variabel *total intl minutes*, mean, dan median yang cenderung sama baik untuk pelanggan dengan status *churn* dan tidak *churn*.

Exploratory Data Analysis

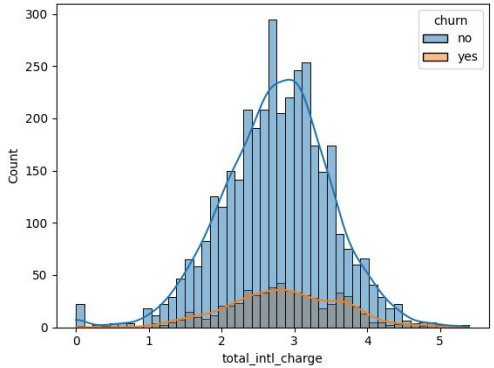
Insight based on Descriptive Statistics

Histogram *Total Intl Calls* vs *Churn*



Secara deskriptif, pelanggan – pelanggan dengan jumlah panggilan internasional yang cenderung rendah akan berpeluang untuk *churn*.

Histogram *Total intl Charge* vs *Churn*

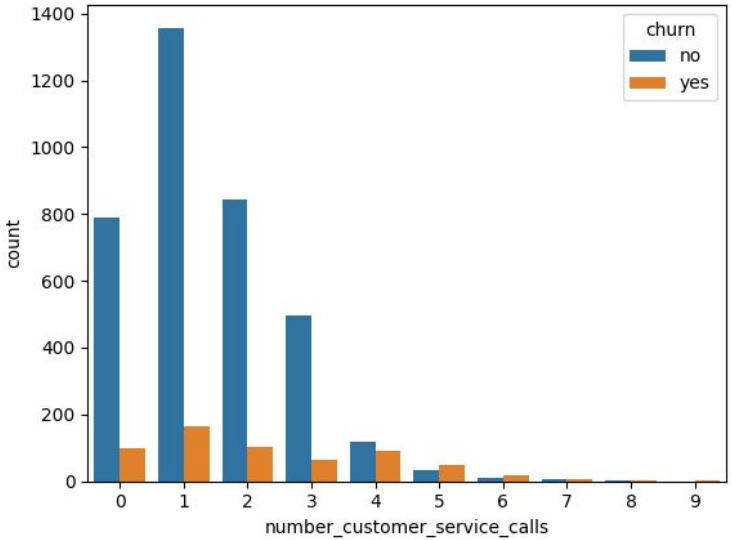


Secara deskriptif, tidak terdapat perbedaan yang signifikan antara pelanggan dengan status *churn* dan tidak *churn* dalam hal biaya panggilan internasional yang dikeluarkan. Hal ini dapat dilihat dari distribusi variabel *total intl charge*, mean, dan median yang cenderung sama baik untuk pelanggan dengan status *churn* dan tidak *churn*.

Exploratory Data Analysis

Insight based on Descriptive Statistics

Histogram *Number Customer Service Calls* vs *Churn*



Secara deskriptif, pelanggan-pelanggan dengan jumlah layanan panggilan yang cenderung banyak akan lebih berpeluang untuk *churn*.

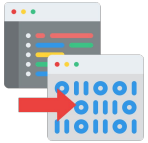
Data Preparation

Data Quality

Tidak ada *missing value* & data duplikat

Data Selection

Drop variabel state



Data Encoding

One hot encoding pada variabel *Area code*, *International plan*, *voice mail plan* & ***Label encoder*** pada variabel *churn*.

Splitting Dataset

Test size = 0.2
Train size = 0.8

Standarization

StandarScaler()

Imbalance Handling

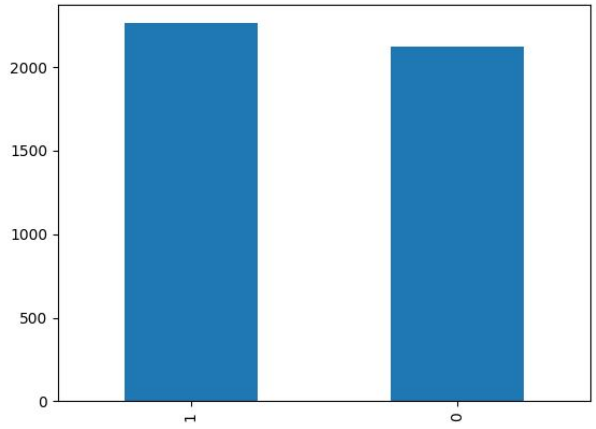
SMOTEENN Hybrid Technique, dengan *sampling strategy* = 0.8

Data Preparation

Imbalance Handling SMOTEENN

SMOTE + ENN adalah teknik hibrid dimana lebih banyak pengamatan yang tidak dihapus dari ruang sampel. Di sini, ENN adalah teknik undersampling lainnya di mana tetangga terdekat dari masing-masing kelas mayoritas diestimasi. Jika tetangga terdekat salah mengklasifikasikan contoh tertentu dari kelas mayoritas, maka contoh tersebut akan dihapus. Mengintegrasikan teknik ini dengan data oversampling yang dilakukan oleh SMOTE membantu dalam melakukan pembersihan data secara ekstensif. Hal ini menghasilkan pemisahan kelas yang lebih jelas dan ringkas (Satpathy, 2023).

Distribusi Kelas Setelah Penanganan Data Imbalance



SMOTEENN dilakukan pada data training dengan sampling strategi = 0.8, hasilnya membuat kelas menjadi sembang. Setelah melalui seluruh tahap *data preprocessing* langkah selanjutnya yaitu tahap pemodelan yang akan dibahas pada slide berikutnya.

Modelling

Classification Algorithm

XGBoost

XGBoost (eXtreme Gradient Boosting) adalah algoritma *machine learning* yang memiliki performa tinggi dan sangat efektif dalam prediksi. Dengan teknik *boosting* dan optimisasi berdasarkan gradien fungsi objektif, XGBoost menghasilkan model yang kuat dan akurat. Selain itu, algoritma ini dapat menangani data yang tidak seimbang dan mengidentifikasi fitur-fitur yang saling berinteraksi. Kelebihan ini menjadikan XGBoost pilihan populer dalam kompetisi *machine learning* dan analisis data.

Gradient Boosting

Gradient Boosting adalah algoritma *machine learning* yang menggabungkan beberapa model lemah menjadi model yang kuat. Kelebihannya terletak pada kemampuannya dalam menangani data kompleks dengan noise atau kesalahan. Dengan mengoptimalkan fungsi objektif berdasarkan gradien pada setiap titik, Gradient Boosting membangun model secara iteratif dan menghasilkan prediksi yang akurat.

Modelling

Why XGBoost and Gradien Boosting can reduce negative impact of Outliers?

XGBoost

XGBoost adalah implementasi populer dari gradient boosting yang dikenal karena kecepatan dan akurasinya. XGBoost juga menyertakan sejumlah fitur yang membuatnya lebih kuat terhadap outlier, seperti kemampuan untuk menentukan pengurangan kerugian minimum untuk setiap pohon (ncrefe, 2023).

Gradient Boosting

Gradient boosting adalah algoritma boosting lain yang dikenal kuat terhadap outlier. Hal ini karena gradient boosting cocok dengan serangkaian pohon keputusan pada data, dan pohon keputusan secara alami kuat terhadap outlier (ncrefe, 2023).

Modelling

Pada tahap ini dilakukan pemodelan baik dengan model base dan model yang telah melalui tahap hyperparameter tuning dan *cross validation*. Hyperparameter tuning dilakukan dengan metode *GridSearch.* Pada hyperparameter tuning, dilakukan 10 *KFold cross validation*. Parameter yang diperoleh pada algoritma XGBoost, yaitu *'colsample_bytree': 0.5, 'learning_rate': 0.15, 'max_depth': 8, 'min_child_weight': 1, 'subsample': 0.8*. Selanjutnya, pada algoritma Gradient Boosting diperoleh parameter: *'learning_rate': 1, 'loss': 'exponential', 'max_depth': 4, 'max_features': 'log2', 'min_samples_leaf': 1, 'n_estimators': 200*.

Selanjutnya, akan dibandingkan performa model pada model base dan model yang telah melalui tahap *hyperparameter tuning* dan *cross validation* untuk menentukan model terbaik yang akan dibahas pada bagian model evaluation.

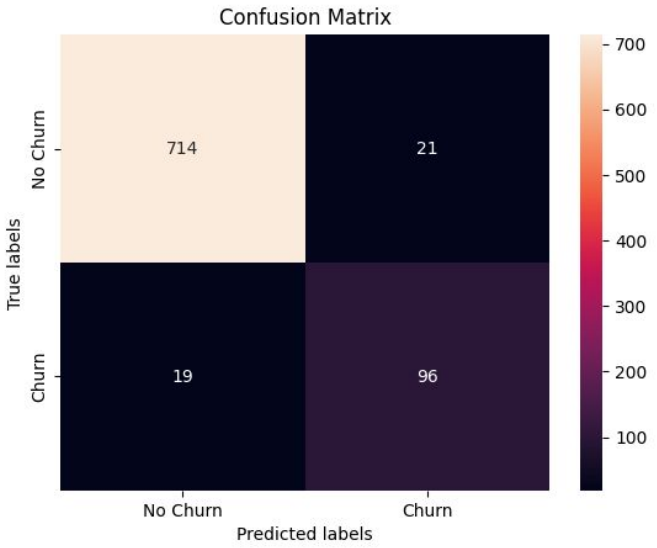
Model Evaluation

Algoritma	Model	Kelas	Precision	Recall	F1-score	Akurasi
XGBoost	Base	0	0.97	0.96	0.97	0.95
		1	0.79	0.83	0.81	
	Tuning	0	0.97	0.97	0.97	0.95
		1	**0.82**	**0.83**	**0.83**	
Gradien	Base	0	0.97	0.94	0.95	0.92
Boosting		1	0.67	0.83	0.74	
	Tuning	0	0.97	0.96	0.96	0.94
		1	**0.76**	**0.81**	**0.78**	

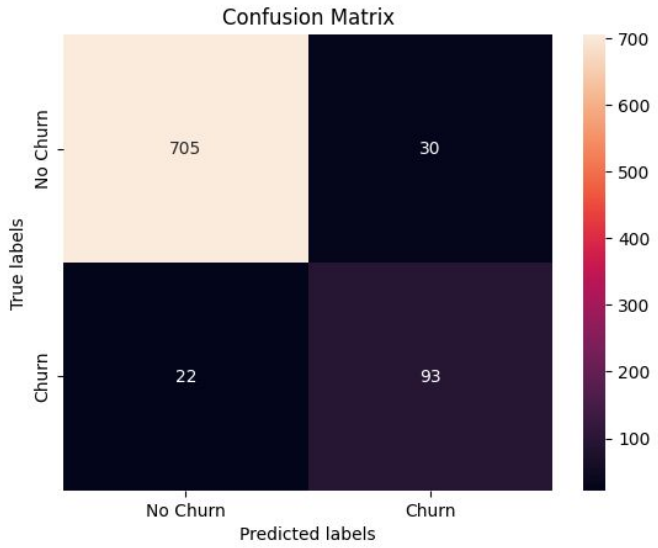
Pemodelan yang melalui tahap *hyperparameter tuning* dan *cross validation* baik pada algoritma XGBoost dan Gradien Boosting memiliki metrik evaluasi yang lebih stabil daripada model base. Selain itu, melalui metrik evaluasi, Presisi, Recall, dan F1-score algoritma **XGBoost memberikan performa model yang lebih baik dibandingkan Gradien Boosting.**

Model Evaluation

Confusion Matrix Algoritma XGBoost dengan Tuning



Confusion Matrix Algoritma Gradien Boosting dengan Tuning



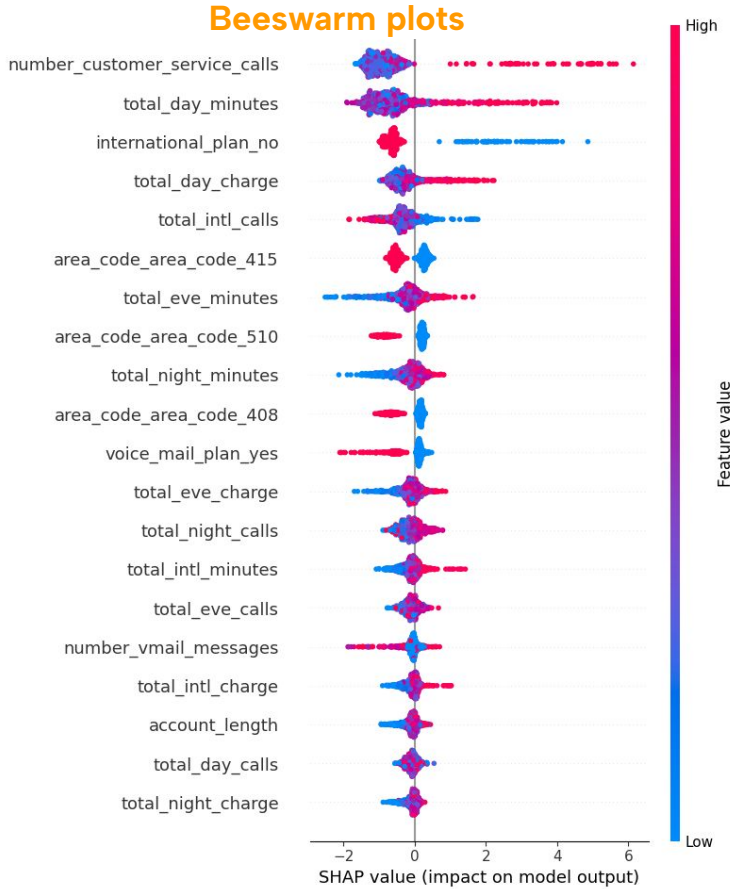
- Biaya mempertahankan pelanggan yang sudah ada jauh lebih rendah daripada mengakuisisi pelanggan baru. Maka dari itu, **meminimkan False Negative** atau kondisi dimana membiarkan pelanggan berhenti berlangganan harus diminimumkan karena dapat menghemat biaya yang dikeluarkan.

Model Evaluation

- Berdasarkan confusion matrix pada slide sebelumnya, dari 850 pelanggan, melalui algoritma XGBoost dan Gradien Bossting secara berurutan terdapat kesalahan klasifikasi pelanggan sebanyak 40 dan 52. Dimana, pada algoritma XGBoost memiliki False Negative yang lebih rendah dari pada Gradien Boosting.
- Maka dari itu, ditinjau dari performa model dan peluang penghematan biaya, dipilih model dengan **Algoritma XGBoost yang telah melalui tahap hyperparameter tuning dan cross validation sebagai model terbaik**. Model ini memiliki parameter `colsample_bytree = 0.5, learning_rate = 0.15, max_depth=8, min_child_weight = 1, subsample=0.8`

Insight & Recommendation

Feature Importance



Top 5 fitur yang paling berpengaruh terhadap customer churn:

1. Number_customer_service _calls
2. Total_day_minutes
3. International_plan_no
4. Total_day_charge
5. Total_intl_calls

Insight based on Feature Importance

1. ***Number customer service calls***

   Pelanggan-pelanggan dengan jumlah layanan panggilan yang cenderung banyak akan lebih berpeluang untuk churn.

2. ***Total days minutes***

   Pelanggan – pelanggan dengan jumlah panggilan siang hari dalam menit yang cenderung tinggi akan berpeluang untuk churn.

3. ***International_plan_no***

   Pelanggan – pelanggan yang memiliki paket panggilan internasional cenderung berpeluang untuk churn.

4. ***Total day charge***

   Pelanggan-pelanggan dengan total biaya panggilan siang hari yang mahal berpeluang untuk churn.

5. ***Total intl calls***

   Pelanggan-pelanggan dengan jumlah panggilan internasional yang cenderung rendah berpeluang untuk churn.

Recommendation

1. Meningkatkan kualitas layanan pelanggan serta memperkenalkan solusi mandiri seperti basis pengetahuan online atau alat bantuan otomatis untuk mengurangi jumlah panggilan layanan yang diperlukan.

2. Melakukan promosi atau diskon khusus untuk paket panggilan internasional, serta memberikan informasi yang lebih jelas dan menarik tentang manfaat yang diperoleh dengan menggunakan paket ini.

3. Memberikan diskon atau penawaran khusus bagi pelanggan yang menggunakan layanan secara intensif di siang hari, serta mengevaluasi kebutuhan dan preferensi pelanggan untuk menyesuaikan paket yang ditawarkan.

Predict Unseen Data

Pada soal disajikan data test dimana merupakan unseen data yang tidak memiliki label. Prediksi status *churn* pada *unseen* data dilakukan dengan menerapkan *data preprocessing* seperti data latih dan menggunakan model terbaik yang diperoleh melalui tahap pelatihan model. Berikut merupakan data yang berisi hasil prediksi status *churn* dari setiap pelanggan pada unseen data, dimana terdapat dua kolom yaitu id (id pelanggan) dan y (status *churn*, dengan 1 : churn dan 0 : tidak churn).

DataChurn_Prediction.csv

Daftar Pustaka

Meanderings, A. T.-T. (2020). *SHAP - What Is Your Model Telling You? Interpret CatBoost Regression and Classification Outputs*. Youtube. https://www.youtube.com/watch?v=ZkIxZ5xlMuI&t=3s

Ncrefe. (2023). *Robust Algorithms to Outliers*. Medium. https://medium.com/@mefeincir/robust-algorithms-to-outliers-c13ebb51494

Satpathy, S. (2023). *SMOTE for Imbalanced Classification with Python*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2020/10/overcoming-class-imbalance-using-smote-techniques/

Link Tugas

https://bit.ly/TelecomChurnPrediction_Coding

Link Dashboard

BINAR

Thank You

A Winner is a Dreamer who Never Gives up

~ Nelson Mandela