

USE CASE DATA SCIENCE DI BISNIS INDUSTRI RETAIL: ANALISIS CUSTOMER CHURN PREDICTION PADA E-COMMERCE

I. Pendahuluan

Di era digital saat ini, e-commerce telah menjadi salah satu industri retail dengan pertumbuhan tercepat. Hal ini karena e-commerce memberikan kemudahan bagi konsumen untuk melakukan pembelian tanpa harus datang ke toko fisik, beberapa contoh e-commerce yang berkembang pesat seperti Shopee, Tokopedia, Lazada, dan lain sebagainya. Sebagai konsekuensi meningkatnya persaingan yang sengit di pasar e-commerce, merupakan tantangan besar bagi perusahaan untuk mempertahankan pelanggan agar tetap loyal dan tidak meninggalkan platform e-commerce atau customer churn. Customer churn adalah istilah yang digunakan untuk menggambarkan situasi ketika pelanggan atau konsumen sebuah perusahaan berhenti menggunakan produk atau layanan yang ditawarkan oleh perusahaan tersebut dan beralih ke pesaing atau tidak menggunakan produk atau layanan tersebut sama sekali. Customer churn merupakan masalah serius bagi banyak perusahaan karena dapat mengakibatkan penurunan pendapatan dan mengganggu pertumbuhan bisnis perusahaan. Oleh karena itu, menganalisis penyebab customer churn dapat membantu perusahaan mengidentifikasi masalah dan menemukan solusi untuk mempertahankan pelanggan.

Untuk mengatasi tantangan ini, penerapan data science untuk menganalisis customer churn menjadi semakin penting dalam menjaga keberlangsungan bisnis e-commerce. Data science dapat membantu bisnis e-commerce untuk memahami perilaku pelanggan, memprediksi kemungkinan pelanggan keluar atau customer churn, dan mengambil tindakan retensi yang sesuai untuk mempertahankan pelanggan. Perusahaan e-commerce dapat memanfaatkan data pelanggan dan melalui analisis data yang mendalam, Perusahaan e-commerce dapat mengidentifikasi faktor-faktor yang mempengaruhi customer churn dan menentukan strategi yang efektif untuk mengurangi tingkat churn tersebut serta mengambil tindakan yang tepat untuk mempertahankan pelanggan yang sudah ada dan menarik pelanggan baru.

Dalam esai ini, akan dibahas lebih lanjut mengenai penerapan data science dalam menganalisis customer churn pada e-commerce. Dalam konteks ini, data science akan diterapkan untuk mengumpulkan dan menganalisis data pelanggan e-commerce dengan tujuan untuk mengembangkan model prediksi yang akurat untuk memprediksi kemungkinan pelanggan beralih atau churn. Churn customer prediction bertujuan untuk memprediksi peluang pelanggan untuk beralih atau churn, sehingga e-commerce dapat memberikan perhatian yang meminimalkan peluang mereka untuk beralih. Prediksi dibuat menggunakan *machine learning*, dengan menggunakan salah satu algoritmanya yaitu regresi logistik, dimana algoritma regresi logistik akan mengklasifikasikan pelanggan sebagai churn di masa mendatang atau tidak, dengan menerapkan metodologi Cross Industry Standard Process for Data Mining (CRISP-DM).

II. Metodologi

Metodologi yang digunakan adalah proses machine learning yang mengacu pada proses data mining oleh the Cross Industry Standard Process for Data Mining (CRISP-DM). CRISP-DM

adalah model proses metodologi data mining, model ini terdiri dari enam fase yang saling berhubungan yaitu:

1. Memahami proses bisnis (business understanding)

Business Understanding adalah langkah pertama dalam metodologi CRISP-DM dan langkah yang sangat penting dalam melakukan analisis data. Pada fase ini, analisis data perlu memahami tujuan bisnis yang akan dicapai dengan analisis data tersebut. Fase ini melibatkan pengumpulan informasi tentang tujuan bisnis, kebutuhan bisnis, dan kendala bisnis yang ada. Tahap Business Understanding membantu memastikan bahwa analisis data yang dilakukan dapat menghasilkan nilai tambah bagi bisnis dan dapat memecahkan masalah yang muncul.

2. Memahami data (data understanding)

Pada fase ini, tujuannya adalah untuk memahami data yang akan digunakan dalam analisis. Hal ini mencakup pemahaman tentang sumber data, kualitas data, dan karakteristik data. Dilakukan dengan mencari, mengumpulkan, dan memperoleh data dari berbagai sumber yang relevan dengan tujuan analisis.

3. Persiapan data (data preparation)

Persiapan data (data preparation) adalah salah satu tahapan penting dalam proses CRISP-DM. Beberapa hal yang dapat dilakukan dalam persiapan data dalam CRISP-DM meliputi:

- Pembersihan data, seperti pembersihan data dari nilai yang hilang, duplikat, dan tidak valid. Hal ini dilakukan agar data yang digunakan dalam analisis lebih akurat.
- Transformasi data, tahapan ini melibatkan transformasi data menjadi format yang sesuai dengan kebutuhan analisis. Contohnya adalah melakukan normalisasi data atau mengubah data kategori menjadi data numerik.
- Integrasi data, tahapan ini melibatkan penggabungan data dari berbagai sumber menjadi satu kesatuan.
- Formatting data, tahapan ini melibatkan format data agar sesuai dengan kebutuhan analisis.
- Pengujian data, tahapan ini melibatkan pengujian data yang sudah dipersiapkan sebelumnya untuk memastikan data tersebut akurat dan dapat digunakan untuk analisis.

4. Pengembangan model (modeling)

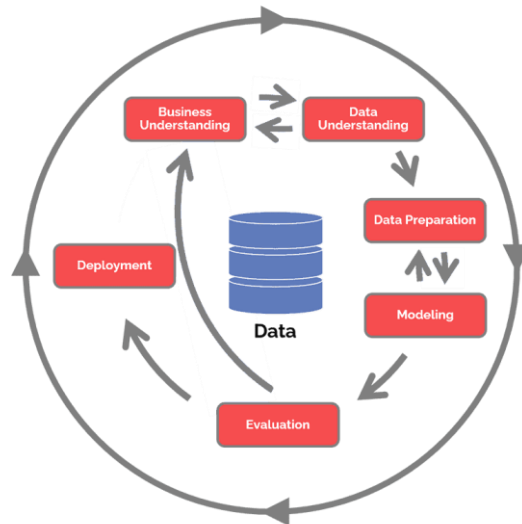
Pada tahap ini, perlu dilakukan pemilihan algoritma yang sesuai untuk mengembangkan model. Ada beberapa metode terkait algoritma machine learning yang dapat digunakan dalam pengembangan model ini, seperti regresi logistik, decision tree, random forest, neural network, dan sebagainya. Pada tahap ini, perlu juga dilakukan pembagian data atau split data dengan membagi data menjadi dua bagian, yaitu data training dan data testing. Data training digunakan untuk mengembangkan model, sedangkan data testing digunakan untuk menguji model. Hal ini dilakukan untuk melakukan validasi model guna memastikan bahwa model yang dikembangkan memiliki akurasi yang tinggi dan dapat diandalkan.

5. Evaluasi (evaluation)

Mengevaluasi model yang telah dibangun, untuk memastikan bahwa model tersebut sesuai dengan masalah bisnis yang ingin diselesaikan, dan memiliki performa yang baik dalam memprediksi atau menjelaskan data. Tahap ini melibatkan pengukuran kinerja dan pengembangan model alternatif jika diperlukan.

6. Implementasi (deployment)

Tahap terakhir melibatkan penerapan model yang telah dikembangkan ke dalam lingkungan produksi atau bisnis. Tahap ini melibatkan integrasi model ke dalam proses bisnis dan pelatihan pengguna untuk menggunakan model.



Gambar Diagram Alir dari CRIPS-DM

Dari beberapa algoritma machine learning yang dapat digunakan terkait pengembangan model prediksi churn customer, pada esai ini akan menggunakan algoritma regresi logistik. Regresi logistik adalah suatu metode statistik yang digunakan untuk memodelkan hubungan antara variabel independen (prediktor) dan variabel dependen biner (target), yaitu variabel yang hanya memiliki dua nilai, misalnya ya atau tidak, 0 atau 1, hidup atau mati, sukses atau gagal, dan sebagainya. Regresi logistik digunakan untuk mengestimasi probabilitas kejadian variabel dependen biner berdasarkan nilai-nilai variabel independen.

Regresi logistik menghasilkan model logistik, yang merupakan persamaan matematis yang menggambarkan hubungan antara variabel independen dan variabel dependen biner. Model logistik ini dapat digunakan untuk memprediksi kemungkinan kejadian variabel dependen biner berdasarkan nilai-nilai variabel independen yang diberikan. Salah satu keuntungan regresi logistik adalah kemampuannya untuk menangani data biner, kategori, dan numerik dalam satu model, sehingga cocok digunakan dalam berbagai aplikasi. Dalam customer churn prediction pada *e-commerce*, regresi logistik akan mengklasifikasikan pelanggan sebagai churn di masa mendatang atau tidak.

Jika terdapat n buah variabel masukan yang disimbolkan $x = (x_1, x_2, x_3, \dots, x_n)$ dan Y adalah variabel dependen dimana peluang $P(Y = 1 | x) = \pi(x)$, maka persamaan regresi logistik dituliskan sebagai berikut:

$$y(x) = \ln \left(\frac{\pi(x)}{1 - \pi(x)} \right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n$$

III. Pembahasan

Adapun penjelasan secara lebih detail mengenai metodologi yang dilakukan dalam use case terkait *customer churn prediction* pada *e-commerce* adalah sebagai berikut:

1. Pemahaman Proses Bisnis (business understanding)

Pada tahap ini, akan terkait pemahaman masalah bisnis yang ingin diselesaikan. Di industri retail seperti *e-commerce*, model *churn prediction* digunakan untuk memprediksi tingkat kemungkinan pelanggan berpindah dari satu *e-commerce* ke *e-commerce* lainnya, dengan demikian *e-commerce* dapat memberikan *reatment* untuk meminimalisir kemungkinan mereka *churn*. Tujuan bisnis dari proyek prediksi churn, yaitu meningkatkan retensi pelanggan atau mengurangi biaya akuisisi pelanggan baru.

2. Pemahaman Data (data understanding)

Langkah selanjutnya adalah memahami data yang tersedia dan memahami karakteristik pelanggan yang ingin diprediksi. Data dapat diperoleh dari database pelanggan, dan dapat berisi informasi seperti riwayat transaksi, preferensi produk, dan profil pelanggan. Pada tahap ini, perlu juga untuk menentukan variabel-variabel yang akan digunakan untuk memprediksi perilaku pelanggan.

Untuk menganalisis Customer Churn di *e-commerce*, beberapa data pelanggan yang dapat dikumpulkan antara lain:

- a. Informasi personal pelanggan seperti nama, alamat email, nomor telepon, usia, jenis kelamin, dan lokasi geografis.
- b. Riwayat pembelian pelanggan termasuk jumlah pembelian, frekuensi pembelian, dan produk yang dibeli.
- c. Riwayat interaksi pelanggan dengan website atau aplikasi *e-commerce* seperti waktu yang dihabiskan di situs, jumlah klik, dan produk yang dilihat.
- d. Preferensi pembelian pelanggan, seperti jenis produk yang disukai, merek, kategori, dan harga.
- e. Rating atau ulasan pelanggan tentang produk yang telah dibeli.
- f. Metode pembayaran yang digunakan pelanggan.
- g. Perilaku pelanggan seperti waktu pembelian, waktu pengiriman, dan preferensi pengiriman.

3. Pemrosesan Data (data preparation)

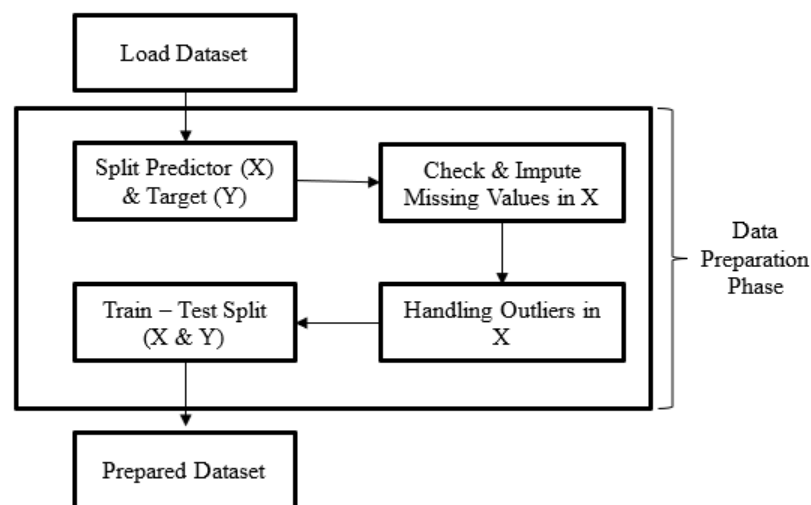
Beberapa hal yang dapat dilakukan dalam persiapan data adalah pemisahan variabel prediktor dan variabel respon, pemeriksaan data hilang yang dapat dilakukan dengan pembersihan data ataupun imputasi, menangani penceilan serta pemisahan data training and testing.

Pemisahan variabel respon dan prediktor, dimana variabel respon yang digunakan adalah variabel churn yang merupakan variabel biner dengan nilai 1 sebagai churn dan 0 tidak churn. Sedangkan variabel lainnya yang dirasa mempengaruhi churn akan digunakan sebagai variabel prediktor, dimana variabel ini nantinya akan dilakukan pengujian untuk memastikan apakah memenuhi asumsi atau tidak. Selain itu, variabel prediktor akan digunakan untuk memprediksi variabel target/respon dengan menggunakan algoritma *machine learning*.

Pemeriksaan data yang hilang, tahap ini dapat dilakukan dengan menghilangkan data yang hilang tersebut atau dapat juga dilakukan dengan imputing. Imputasi adalah metode pengisian data hilang/missing value dengan mengganti data hilang tersebut dengan suatu nilai yang diperoleh dari perhitungan statistik seperti nilai rata-rata ataupun median, hal ini dilakukan agar tidak ada data atau informasi yang terbuang (akibat penghapusan baris) dan hasil analisis yang dihasilkan tetap valid.

Penanganan pencilan adalah penggantian nilai suatu objek yang dianggap menyimpang dari nilai-nilai lainnya pada suatu variabel, sehingga bisa menyebabkan bias. Salah satu cara penanganan pencilan adalah dengan menetapkan standar nilai 3 sigma, dimana amatan yang berada di luar batas $\mu+3\sigma$ (batas atas) atau $\mu-3\sigma$ (batas bawah), maka amatan tersebut harus digantikan dengan nilai sesuai batas atas atau batas bawah. Dengan begitu, efek bias akibat adanya pencilan dapat diminimalisir.

Pemisahan data train-test dilakukan dengan membagi secara acak dataset awal menjadi dua bagian dengan proporsi yang berbeda, umumnya data train memiliki proporsi yang lebih besar dibanding data test. Data training digunakan untuk mengembangkan model, sedangkan data testing digunakan untuk menguji model. Hal ini dilakukan untuk melakukan validasi model guna memastikan bahwa model yang dikembangkan memiliki akurasi yang tinggi dan dapat diandalkan. Beberapa proporsi yang lazim digunakan adalah 90%-10%, 80%-20% dan 70% - 30%.

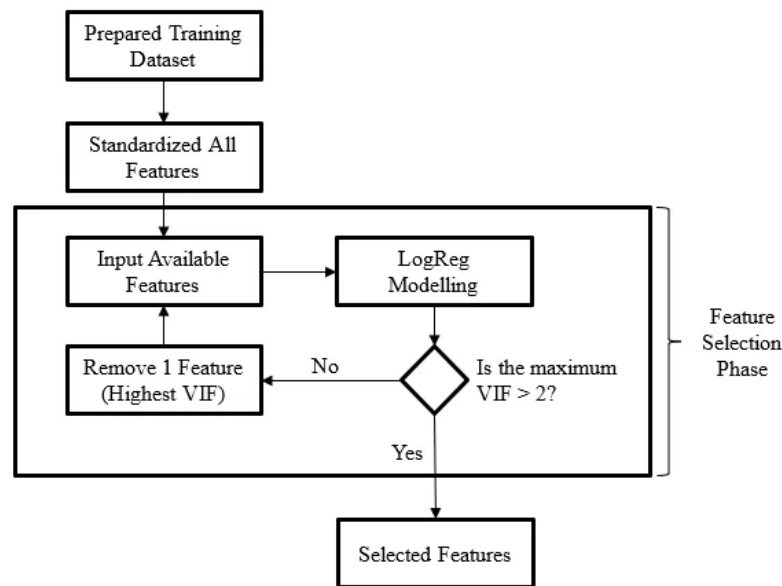


Gambar tahap pemrosesan data

4. Pengembangan Model (Modelling)

Sebelum melakukan pengembangan model, akan dilakukan terlebih dahulu tahap feature engineering & elimination phase. feature engineering adalah tahap mengubah nilai suatu variabel/fitur menjadi suatu nilai yang baru yang dapat dilakukan dengan standarisasi (*standard scaler*), standarisasi sendiri digunakan ketika variabel yang digunakan dalam modeling memiliki satuan yang berbeda, dimana apabila data yang digunakan dengan satuan yang berbeda tetap dilakukan untuk pemodelan akan berdampak pada akurasi model yang dihasilkan. Standarisasi dilakukan dengan dengan

mengubah semua nilai variabel menjadi sekumpulan nilai yang menyebar normal baku dengan rata-rata (μ) = 0 dan standar deviasi (σ) = 1. Sedangkan feature selection adalah proses mereduksi variabel/fitur. Feature selection dilakukan dengan menggunakan *Variance Inflation Factors* (VIF), yaitu sebuah nilai yang mengukur seberapa besar peningkatan varian dari koefisien regresi ketika antar variabel/fitur saling berkorelasi. Proses pengukuran VIF dilakukan secara *looping* dengan metode *backward regression*, dimana semua variabel/fitur digunakan di tahap awal, lalu variabel/fitur dengan nilai VIF tertinggi akan dieliminasi. Proses *looping* berhenti ketika semua peubah/fitur memiliki nilai VIF di bawah nilai *threshold* yaitu 2, 5, atau 10. Alternatif lain untuk eliminasi variabel/fitur diantaranya *Correlation Matrix* atau *Feature Importance*.



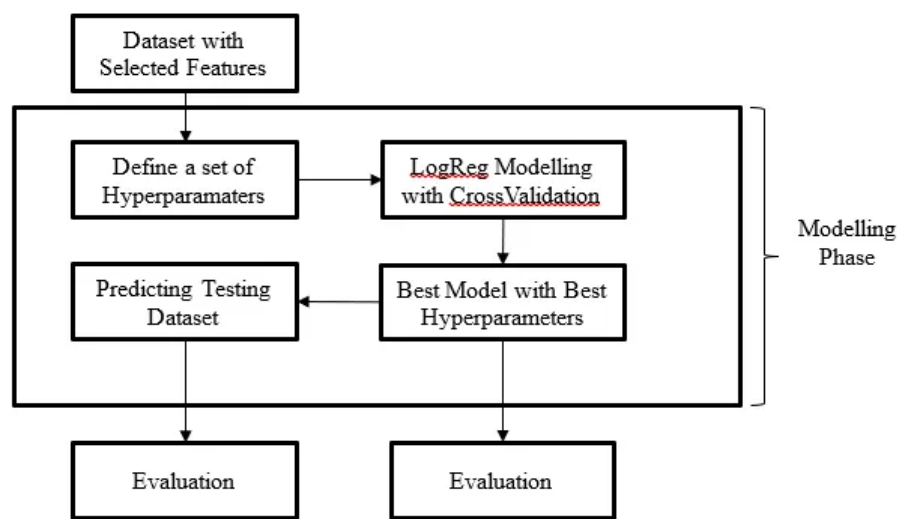
Gambar tahap feature engineering & elimination phase

Tahap pengembangan model adalah tahap membangun model prediksi dengan menggunakan data training, yang selanjutnya akan dilakukan validasi model dengan menggunakan data testing untuk memastikan bahwa model yang dikembangkan memiliki akurasi yang tinggi dan dapat diandalkan.

Pembangunan model menggunakan data training bertujuan untuk menemukan hyperparameter terbaik, yaitu nilai-nilai yang ditetapkan sebelum proses pembangunan model. Proses penentuan hyperparameter dilakukan menggunakan metode randomized cross-validation karena bersifat trial and error, yang dilakukan dengan memilih kombinasi hyperparameter secara acak dan dilanjutkan dengan *k-fold cross validation* untuk menentukan nilai akurasi dari kombinasi hyperparameter tersebut. Proses pemilihan hyperparameter dilakukan sebanyak n kali, sehingga nantinya akan dipilih kombinasi hyperparameter yang mampu menghasilkan akurasi yang tinggi dan dapat diandalkan. Berikut adalah tabel rentang nilai yang dapat digunakan dalam menetapkan nilai hyperparameter pada regresi logistik.

Hyperparameter	Value	Description
Penalty	[11,12]	Type of regularization
C	Uniform (0,4)	Inverse of regularization strength

Sedangkan validasi model yang dilakukan dengan menggunakan data testing bertujuan untuk mencegah terjadinya over-fitting, yaitu kondisi dimana model hanya memberikan hasil prediksi terbaik pada data training, namun memberikan hasil prediksi yang buruk jika memprediksi data testing atau data riil di masa mendatang. Model yang baik adalah model yang mampu memberikan performa bagus di fase training dan testing dengan memberikan hasil prediksi dengan akurasi yang tidak jauh berbeda antara data training dan data testing, sehingga model tersebut dapat diandalkan untuk memprediksi data-data baru di masa mendatang.



Gambar Tahap Pengembangan Model

5. Evaluasi (Evaluation)

Salah satu metrik evaluasi yang umum digunakan dalam mengukur kemampuan model adalah *confusion matrix* yang terdiri dari True Positive (TP), True Negative (TN), False Positive (FP), dan False Negative (FN). Detail dari confusion matrix dapat dilihat dari tabel berikut.

Aktual/Prediksi	Positif	Negatif
Positif	True Positive (TP)	False Negative (FN)
Negatif	False Positive (FP)	True Negative (TN)

Keterangan:

- True Positive (TP) : Jumlah instance kelas positif yang benar diprediksi sebagai kelas positif
- False Positive (FP) : Jumlah instance kelas negatif yang diprediksi sebagai kelas positif
- False Negative (FN) : Jumlah instance kelas positif yang diprediksi sebagai kelas negatif

- True Negative (TN) : Jumlah instance kelas negatif yang benar diprediksi sebagai kelas negatif

Beberapa perhitungan untuk mengevaluasi hasil dapat diketahui dengan matriks evaluasi yaitu, recall, precision, F1-Measure dan akurasi.

a. Recall (Sensitivity)

Recall dihitung untuk mengevaluasi seberapa besar coverage suatu model dalam memprediksi suatu kelas tertentu. Recall didapatkan dengan menghitung perbandingan antara jumlah data untuk satu kelas tertentu yang diprediksi dengan benar dibagi jumlah total kelas tersebut.

$$Recall = \frac{TP}{TP + FN}$$

b. Precision (Presisi)

Precision dihitung untuk mengevaluasi seberapa baik kepastian model dalam memprediksi suatu kelas dengan benar. Precision merupakan perhitungan untuk mendapat nilai antara perbandingan antara jumlah data untuk satu kelas tertentu yang diprediksi dengan benar dibagi jumlah keseluruhan prediksi kelas.

$$Precision = \frac{TP}{TP + FP}$$

c. F1-Measure

F1-Measure adalah perhitungan kombinasi antara recall dan precision. Rentang nilai precision adalah 0 sampai 1, jika nilai mendekati 0 maka model prediksi tidak baik dan mendekati 1 jika model prediksi baik. Untuk mendapatkan nilai dalam persentase maka nilai dikalikan dengan 100.

$$F1 - Measure = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

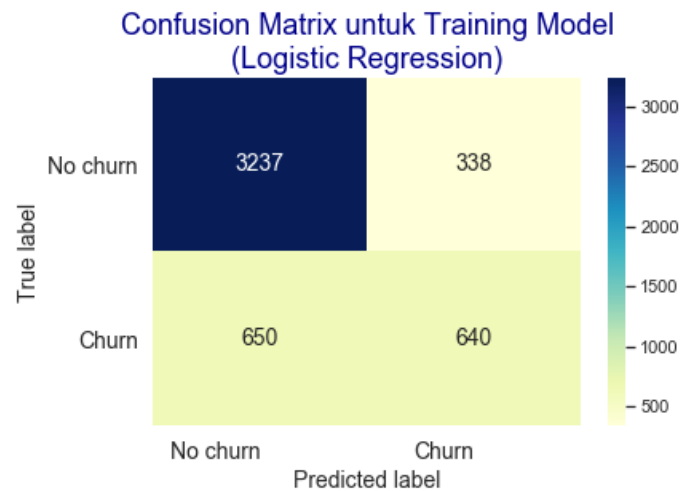
d. Akurasi

Akurasi adalah nilai ketepatan model memprediksi data dengan perbandingan data aktualnya dan sebagai pengukur model seberapa akurat untuk melakukan prediksi.

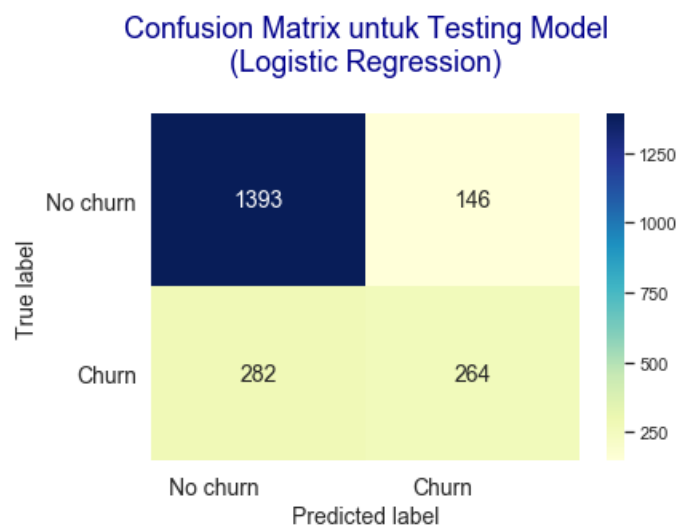
$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP}$$

Semakin tinggi nilai keempat ukuran tersebut maka semakin baik model yang telah dihasilkan. Akurasi adalah metrik evaluasi yang paling sering digunakan untuk menentukan baik tidaknya suatu model, namun akurasi sangat rentan menimbulkan bias ketika menghadapi data dengan kelas yang memiliki ketimpangan yang sangat tinggi. Pada data tersebut, akurasi yang tinggi bisa jadi disebabkan oleh model yang hanya mampu memprediksi kelas mayoritas namun gagal memprediksi kelas minoritas. Ketika jumlah suatu kelas sangat dominan dibanding kelas lainnya, maka akurasi tidak bisa dijadikan tolak ukur kebaikan suatu model sehingga ukuran lain juga harus disertakan seperti sensitivitas atau presisi. Terdapat kasus yang menunjukkan nilai akurasi tinggi namun buruk terhadap prediksi churn, yaitu ketika jumlah data non-churn memiliki ketimpangan yang sangat tinggi. Pengukuran yang lebih tepat dalam model prediksi tanpa imbalance data adalah F1-Measure.

Berikut adalah contoh dari visualisasi metrik evaluasi *confusion matrix* pada data training dan testing:



Dari data training dapat dilihat bahwa model mampu memprediksi data dengan tingkat akurasi sebesar 80%, dengan detail prediksi Churn yang sebenarnya Churn adalah 640, prediksi tidak Churn yang sebenarnya tidak Churn adalah 3237, prediksi tidak Churn yang sebenarnya Churn adalah 650 dan prediksi churn yang sebenarnya tidak churn adalah 338.



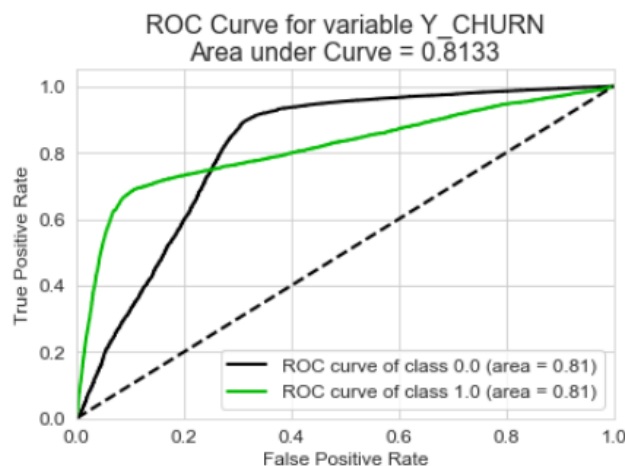
Dari data testing dapat dilihat bahwa model mampu memprediksi data dengan tingkat akurasi sebesar 79%, dengan detail prediksi Churn yang sebenarnya churn adalah 264, prediksi tidak Churn yang sebenarnya tidak Churn adalah 1393, prediksi tidak churn yang sebenarnya Churn adalah 282 dan prediksi Churn yang sebenarnya tidak Churn adalah 146.

Metrik lain yang dapat digunakan adalah kurva *Receiver Operating Character* (ROC), yaitu kurva yang menggambarkan *trade off* antara True Positive Rate (TP/TP+FN) dan False Positf Rate (FP/TN+FP). Semakin cembung kurva tersebut (titik cembung mendekati koordinat 0,1), maka semakin baik model tersebut. Area di bawah kurva

ROC (AUC) adalah ukuran yang berguna yang mengindikasikan kemampuan prediksi pengklasifikasi.

AUC	Description
0.90 - 1.00	Excellent Classification
0.80 - 0.90	Good Classification
0.70 - 0.80	Fair Classification
0.60 - 0.70	Poor Classification
< 0.60	Failure

Contoh dari kurva ROC dapat dilihat pada gambar berikut.



Jika dilihat berdasarkan grafik tersebut diperoleh nilai AUC=0,8133 yang berarti kemampuan klasifikasi prediksi dalam kategori baik.

6. Implementasi (Deployment)

Jika model terbaik telah memenuhi kriteria evaluasi yang sudah ditetapkan, maka model siap untuk digunakan atau di-deploy. Output dari model yang dihasilkan adalah prediksi label per amatan (*not-churn*/ $y=0$ atau *churn*/ $y=1$) serta *probability* setiap amatan untuk *churn* ($y=1$). *Default threshold* yang umum digunakan adalah 0.50, dimana ketika amatan memiliki nilai *probability* > 0.50 , maka akan diberi label prediksi $y=1$, demikian juga sebaliknya.

IV. Kesimpulan

Dalam era digital seperti saat ini, bisnis e-commerce telah berkembang pesat dan menjadi salah satu sektor bisnis yang paling diminati. Namun, bersamaan dengan pertumbuhan bisnis e-commerce yang pesat, juga muncul tantangan baru seperti kehilangan pelanggan atau yang dikenal sebagai customer churn. Customer churn dapat terjadi karena berbagai alasan, termasuk ketidakpuasan pelanggan, persaingan yang ketat, atau perubahan preferensi pelanggan, kualitas produk atau layanan, harga, pengalaman pelanggan, dan faktor lain yang relevan. Oleh karena itu, penting bagi bisnis e-commerce untuk memahami dan menganalisis customer churn agar dapat mengambil tindakan yang tepat untuk mempertahankan pelanggan mereka. Salah satunya adalah dengan menerapkan data science, data science sangat penting dalam menganalisis customer churn karena dapat membantu perusahaan untuk memahami dan

mengidentifikasi faktor-faktor yang mempengaruhi churn, sehingga dapat mengambil tindakan yang tepat untuk mencegah kehilangan pelanggan. Selain itu, data science juga dapat membantu dalam membuat model prediksi churn yang akurat. Dengan menggunakan data historis dan algoritma machine learning, perusahaan dapat membangun model prediksi yang dapat memberikan perkiraan kapan pelanggan akan meninggalkan produk atau layanan, sehingga perusahaan dapat mengambil tindakan untuk mempertahankan pelanggan sebelum terlambat. Dengan demikian, penggunaan data science dalam menganalisis customer churn dapat membantu perusahaan untuk meningkatkan retensi pelanggan, mengurangi biaya akuisisi pelanggan baru, dan meningkatkan keuntungan jangka panjang.

Referensi:

Hanif, Iqbal. (2019). "Implementing Extreme Gradient Boosting (XGBoost) Classifier to Improve Customer Churn Prediction". Proceedings of the 1st International Conference on Statistics and Analytics, ICSA 2019, 2–3 August 2019, Bogor, Indonesia

Mujaddid, Muhammad Faruq. (2017). Analisis Churn Prediction Menggunakan Metode Logistic Regression Dan Smote (Synthetic Minority Over-Sampling Technique) Pada Perusahaan Telekomunikasi. *e-Proceeding of Engineering*. Vol.4, No.3, 2017, hal:5046-5054

Ma'ruf, Hafiz. (2020). Penerapan Data Science pada Marketing (Customer Churn Prediction-Python), diakses pada tanggal 19 Maret 2023, dari <https://hafizmrf3.medium.com/penerapan-data-science-pada-marketing-customer-churn-prediction-python-de43eed149a4>