



Hasil Interpretasi Model Prediksi Customer Churn

Sandarma Natapaima Nainggolan (DS7-17)

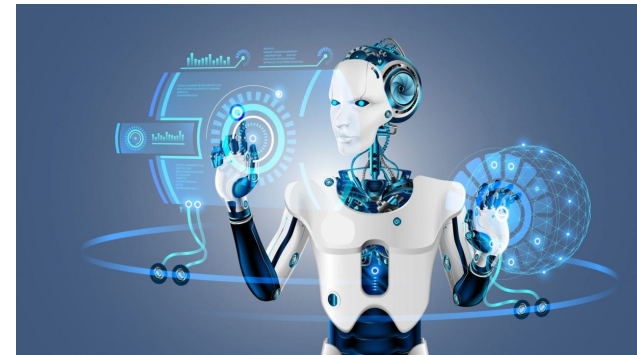
Tahapan

EDA (Exploratory Data Analysis)

- Data Information
- Target Proportion
- Demographic Characteristics
- Distribution

Modelling

- Feature Engineering
- Normalization, Sampling
- Model Selection
- Build Model & Predict Data Test





EDA (Exploratory Data Analysis)

Data Information

- Berapa Shape Data ?
- Apa Saja Variabel Kategori ?
- berapa jumlah variabel numerik ?
- apakah ada missing values?

Target Proportion & Demographics Char

- berapa proporsi target ?
- apakah target proportion nya balance ?
- variabel kategorik mana yang paling mempengaruhi target ?

Penyelesaian

Data Information

- Shape data = (4250,21) ada 4250 baris dan 21 kolom
- variabel kategorik ialah : state, area_code, international_plan dan voice_mail_plan
- jumlah variabel numerik ialah : 14
- Tidak terdapat missing value

Target Proportion & Demographics Char

- 3652 untuk kelas 'no' dan 598 untuk kelas 'yes'
- tidak , percentage nya ialah 85% untuk kelas 'no'
- variabel international plan



Modelling

Feature Engineering & Preprocessing

- Apa metode encoding yang digunakan untuk variabel fitur ?
- Bagaimana cara mengatasi data yang tidak balance ?
- Metode normalisasi apa yang digunakan pada data ?

Model Selection

- Berapa Model yang akan di seleksi ?
- Model apa saja yang akurasi nya bagus ?
- Model apa yang dipilih dan alasan nya apa ?

Penyelesaian

Data Information

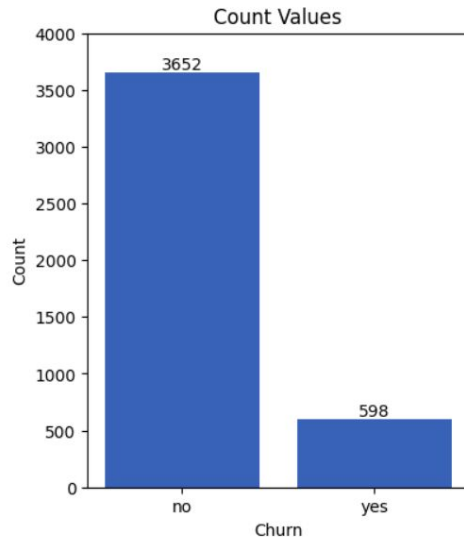
- Menggunakan metode encoding Target Mean Encoding
- Menggunakan metode sampling, yaitu SMOTE, nantinya data sintesis akan dibuat untuk mengisi data kelas minoritas
- Menggunakan metode normalisasi Min Max

Model Selection

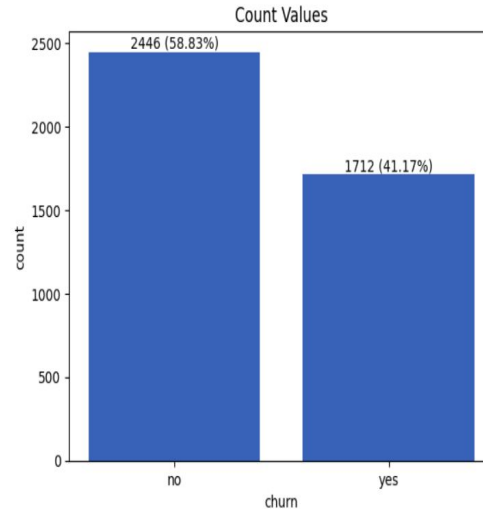
- 7 Model yang akan diseleksi
- Gradient Boosting, XGBoost, dan Random Forest
- Gradient Boosting , karena menghasilkan skor metric yang konsisten dan bagus, baik melalui metode cross-validation maupun dengan metode train-test split.

Resolve data imbalance

Before :



After :




Menggunakan Metode Sampling (SMOTE) :

- Jumlah data kelas mayoritas tetap dipertahankan, sementara jumlah data kelas minoritas ditambahkan dengan data sintesis, sehingga informasi dari data tetap dapat dipertahankan.
- Dalam SMOTE, sampel acuan dari kelas minoritas dipilih. Sampel ini akan menjadi dasar untuk menciptakan sampel sintesis baru.



Model Selection (Cross Validation)

	accuracy	precision_weighted	recall_weighted	f1_weighted	roc_auc	neg_log_loss
Logistic Regression	0.787156	0.786194	0.787156	0.786309	0.859123	-0.466861
K-Nearest Neighbor	0.827800	0.847299	0.827800	0.829107	0.929485	-1.033441
Naive Bayes	0.805678	0.807364	0.805678	0.806232	0.851355	-0.590787
Random Forest	0.965609	0.965545	0.964888	0.962863	0.989817	-0.165650
Decision Tree	0.927854	0.928627	0.927374	0.927912	0.927149	-2.626442
Gradient Boosting	0.955509	0.955293	0.955028	0.955117	0.980941	-0.161439
XGBoost	0.970180	0.970333	0.970180	0.970114	0.985897	-0.114775



Model Selection (Train Test Split)

	accuracy_train	accuracy_val	precision_train	precision_val	recall_train	recall_val	f1_train	f1_val	auc_train	auc_val	log_loss_train	log_loss_val
Logistic Regression	0.795533	0.770032	0.762402	0.731161	0.731219	0.698444	0.746485	0.714428	0.864988	0.845299	4.593434e-01	0.482276
K-Nearest Neighbor	0.893471	0.835737	0.813559	0.746411	0.961603	0.910506	0.881408	0.820333	0.976469	0.923777	2.289344e-01	1.088616
Naive Bayes	0.813746	0.794872	0.759494	0.733696	0.801336	0.787938	0.779854	0.759850	0.856799	0.843282	5.750648e-01	0.621867
Random Forest	1.000000	0.954327	1.000000	0.965377	1.000000	0.922179	1.000000	0.943284	1.000000	0.982704	5.024739e-02	0.180885
Decision Tree	1.000000	0.935096	1.000000	0.927022	1.000000	0.914397	1.000000	0.920666	1.000000	0.931994	2.220446e-16	2.339372
Gradient Boosting	0.975601	0.944712	0.982036	0.945892	0.958264	0.918288	0.970004	0.931885	0.992704	0.976333	1.235604e-01	0.177665
XGBoost	1.000000	0.961538	1.000000	0.969758	1.000000	0.935798	1.000000	0.952475	1.000000	0.984319	7.272123e-03	0.132459

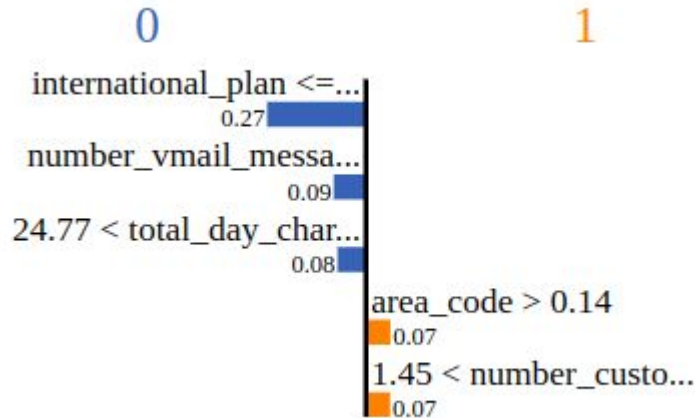


Model Selection (Conclusion)

- Logistic Regression: Model memiliki akurasi yang cukup baik baik pada data training maupun data validasi. Namun, presisi dan recall sedikit lebih rendah, menunjukkan kemungkinan adanya kesulitan dalam mengklasifikasikan kelas positif.
- K-Nearest Neighbor (KNN): KNN memiliki akurasi yang tinggi baik pada data training maupun data validasi. Namun, presisi pada data validasi lebih rendah dibandingkan dengan data training, yang menunjukkan adanya sedikit overfitting. Namun, recall yang tinggi menunjukkan kemampuan model dalam mengidentifikasi sebagian besar kelas positif.
- Naive Bayes: Model Naive Bayes memiliki kinerja yang cukup baik dengan akurasi, presisi, dan recall yang seimbang pada kedua set data, baik training maupun validasi.
- Random Forest dan Decision Tree, XGBoost: Ketiga model ini menunjukkan performa yang sangat baik pada data training, bahkan mencapai akurasi 100%. Namun, kemungkinan besar terjadi overfitting karena terdapat perbedaan kinerja yang signifikan antara data training dan validasi.
- Gradient Boosting: Model ini juga menunjukkan kinerja yang sangat baik, bahkan mendekati akurasi 100% pada data training. Namun, performa yang cukup baik juga terjadi pada data validasi, menunjukkan kemampuan model dalam generalisasi.

Dari kesimpulan di atas, Gradient Boosting mungkin merupakan pilihan yang paling baik karena memiliki kinerja yang sangat baik pada data training, dan mampu menggeneralisasi dengan baik pada data validasi.

Model Interpretation (for class 0)



Predict : 0 , Actual : 0

- **international_plan <= 0.11**: Pelanggan yang tidak memiliki paket rencana internasional memiliki kemungkinan churn yang lebih rendah (nilai koefisien negatif), yang mengindikasikan bahwa pelanggan yang tidak memiliki paket ini cenderung lebih stabil dan cenderung tidak pindah.
- **number_vmail_messages <= 0.00**: Pelanggan yang tidak menggunakan layanan pesan suara (voicemail) memiliki kemungkinan churn yang lebih rendah. Hal ini mungkin menunjukkan bahwa pelanggan yang tidak menggunakan layanan ini cenderung kurang bergantung pada layanan telepon dan oleh karena itu kurang mungkin untuk pindah.

Model Interpretation (For Class 1)



Predict : 1 , Actual : 1

- **international_plan > 0.11**: Pelanggan yang memiliki paket rencana internasional memiliki kemungkinan churn yang lebih tinggi (nilai koefisien positif). Hal ini menunjukkan bahwa pelanggan yang memiliki paket ini cenderung lebih mungkin untuk pindah.
- **number_customer_service_calls > 3.00**: Pelanggan yang melakukan panggilan layanan pelanggan lebih dari 3 kali memiliki kemungkinan churn yang lebih tinggi. Ini menunjukkan bahwa pelanggan yang sering menghubungi layanan pelanggan mungkin mengalami masalah atau ketidakpuasan yang signifikan, yang dapat menyebabkan mereka untuk pindah.