To Churn or Not to Churn: This is the Analysis

Deta Putri Prakoso 1706047315

# Background

Churn didefinisikan sedikit berbeda oleh setiap organisasi atau produk. Umumnya, pelanggan yang berhenti menggunakan produk atau layanan untuk jangka waktu tertentu disebut sebagai churners. Akibatnya, churn adalah salah satu elemen terpenting dalam Key Performance Indicator (KPI) suatu produk atau layanan. Analisis siklus hidup pelanggan yang lengkap memerlukan melihat tingkat retensi untuk lebih memahami kesehatan bisnis atau produk.

Terjadinya churn mahal untuk bisnis. Tingkat churn yang tinggi memaksa bisnis untuk bersaing dengan tekanan dan kesulitan membawa cukup banyak pelanggan baru, guna menutup kerugian akibat churn. Bahkan peningkatan tingkat churn (%) yang tampak kecil dapat dengan cepat memiliki efek negatif besar pada kemampuan perusahaan untuk tumbuh. Terlebih lagi, tingkat churn yang tinggi lebih cenderung bertambah seiring waktu. Itu sebabnya memiliki pemahaman analisis churn sangat penting. Memahami berbagai alasan di balik churn pelanggan adalah langkah mendasar dalam menangani dan mengurangi tingkat churn pelanggan.

Dari perspektif data mining/machine learning, churn dapat dirumuskan sebagai masalah klasifikasi biner. Dalam sebagian besar masalah churn, jumlah churner jauh lebih sedikit dibandingkan jumlah pengguna yang terus bertahan. Ini menyebabkan dataset tidak seimbang dalam jumlah sampel dari setiap kasus.

# Data and Methods

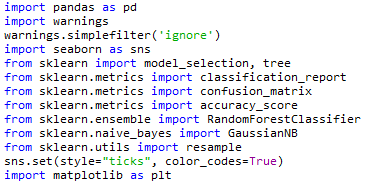
Dataset dapat diunduh pada tautan berikut: <https://data.world/earino/churn> (terlampir juga di email)

Dataset ini dimodifikasi dari yang disimpan di repositori data UCI (yaitu, kode area dan nomor telepon telah dihapus). Ini adalah data buatan yang serupa dengan apa yang ditemukan di profil pelanggan aktual. Dataset ini berisi variabel yang berbeda yang menjelaskan atribut industri telekomunikasi dan berbagai faktor yang dianggap penting ketika berhadapan dengan pelanggan industri telekomunikasi. Variabel target di sini adalah churn yang menjelaskan apakah pelanggan akan churn atau tidak. Kita dapat menggunakan kumpulan data ini untuk memprediksi pelanggan yang akan churn atau yang tidak akan churn tergantung pada berbagai variabel yang tersedia.

|  |  |  |
| --- | --- | --- |
| Response Variable |  |  |
| churn | string | Whether the customer churned or not |
| Explanatory Variables |  |  |
| accountlength | integer | the number of days the user has this account |
| internationalplan | string | Whether the customer has international plan or not |
| voicemailplan | string | Whether the customer has voice mail plan or not |
| numbervmailmessages | integer | Number of voice mail messages |
| totaldayminutes | double | Total minutes of calls during the day |
| totaldaycalls | integer | Number of calls during the day |
| totaldaycharge | double | Total charges of calls during the day in USD |
| totaleveminutes | double | Total minutes of calls during the evening |
| totalevecalls | integer | Number of calls during the evening |
| totalevecharge | double | Total charges of calls during the evening in USD |
| totalnightminutes | double | Total minutes of calls during the night |
| totalnightcalls | integer | Number of calls during the night |
| totalnightcharge | double | Total charges of calls during the night in USD |
| totalintlminutes | double | Total minutes of international calls |
| totalintlcalls | integer | Number of international calls |
| totalintlcharge | double | Total charges of international calls in USD |
| numbercustomerservicecalls | integer | Number of customer service calls |

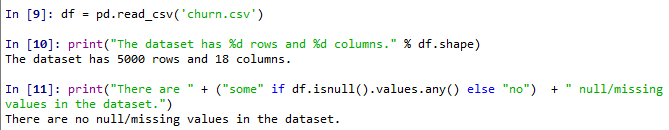
Karena variabel respons berupa binary, pilihan metode klasifikasi yang digunakan kali ini adalah: Decision Tree, Random Forest, dan Naïve Bayes. Metode ini saya pilih karena cepat, tanpa iterasi, robust terhadap outliers dan missing values, serta lebih mudah untuk diinterpretasikan. Pada akhir modeling data, akan dibandingkan metode mana yang lebih baik, dan variabel indepen mana yang paling berpengaruh untuk churn pelanggan.

Untuk data processing dan modelling tersebut, akan digunakan bahasa pemograman Python dengan software Anaconda Spyder; dengan module yang dipakai sbb:



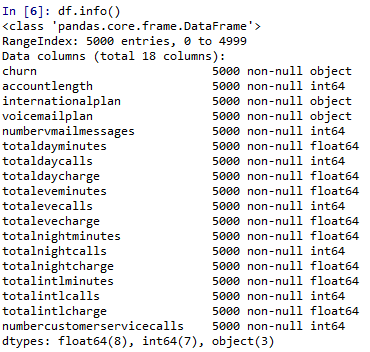
# Data Processing

1. Data Pre-processing

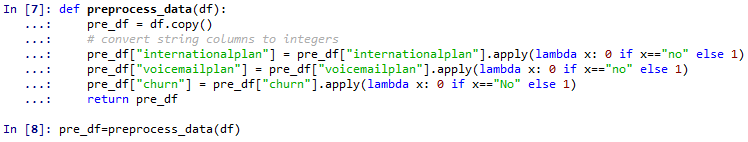
Pertama, kita akan load dataset ke Python environment dan mengecek missing values.

Misalkan terdapat missing value, bisa diabaikan karena model DT dan RF didasarkan pada frekuensi, sehingga adanya missing value ataupun outliers tidak akan membuat model menjadi sensitif.

Selanjutnya kita akan lihat tipe variabel dalam dataset ini.

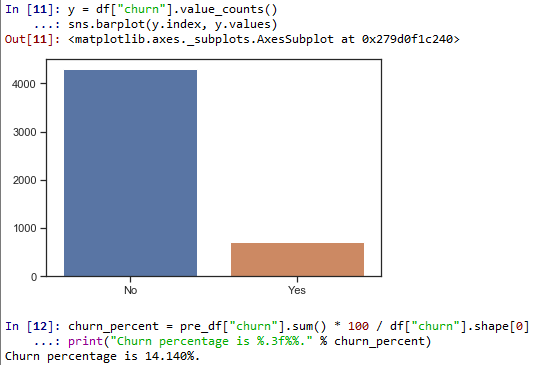


Variabel kategorik churn, internationalplan, dan voicemailplan perlu diubah menjadi variabel dummy untuk processing data yang lebih mudah.



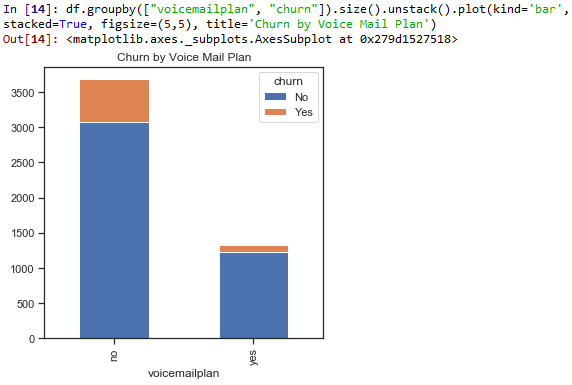
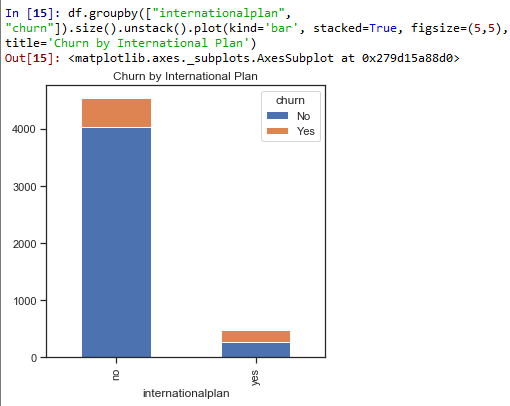
2. Descriptive Statistics

Selanjutnya, kita lihat penyebaran variabel target kita.



Kita dapat melihat bahwa kita memiliki sampel yang jauh lebih banyak untuk pelanggan tidak churn daripada untuk pelanggan dengan churn. Jadi kita memiliki ketidakseimbangan kelas untuk variabel target yang dapat mengarah pada model prediksi yang bias terhadap mayoritas (dalam kasus ini: tidak ada churn). Untuk mengatasi masalah ini, pada bagian berikutnya akan digunakan oversampling saat membangun model.

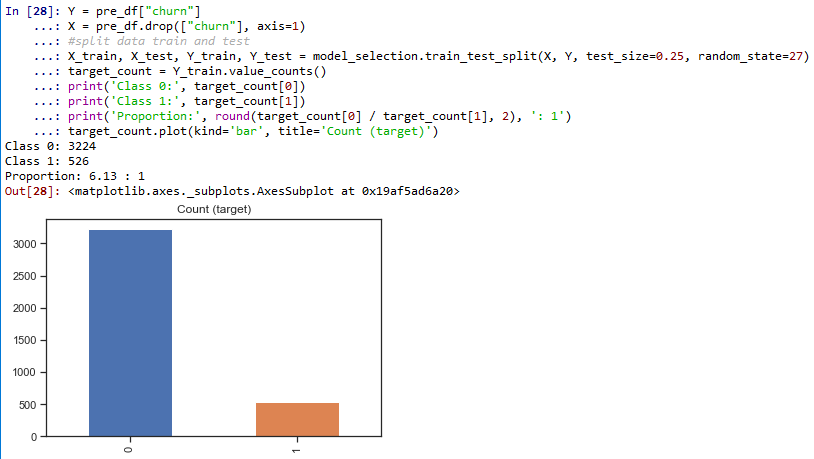
Selanjutnya, kita akan lihat penyebaran churn per variabel independen kategorik



Dari kedua plot di atas dapat kita curigai bahwa pelanggan yang tidak memiliki international plan dan voice mail plan, lebih banyak churn.

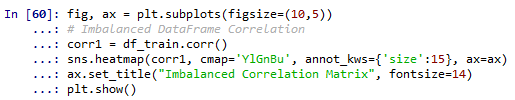
3. Splitting the Data (with no treatment as imbalanced data)

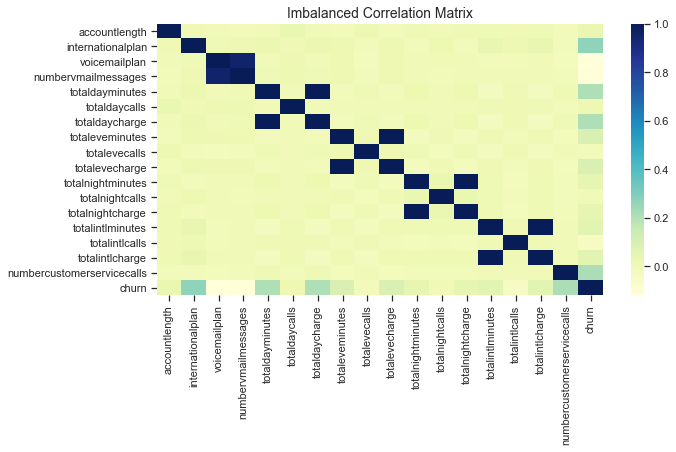
Sekarang dataset akan dibagi menjadi train data dan test data dan dimodelkan TANPA memperlakukannya sebagai imbalanced data.



Bisa dilihat bahwa proporsi no churn : churn = 6 : 1 🡪 very unbalanced

Akan kita lhat korelasi antar variabelnya untuk nantinya dibandingkan setelah oversampling.

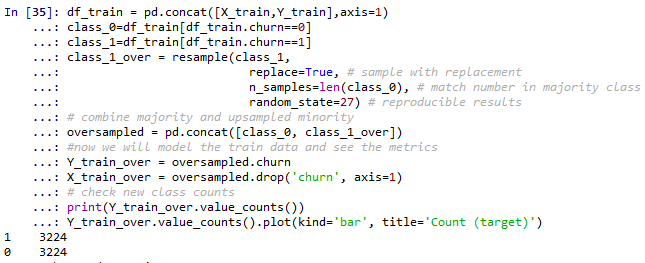


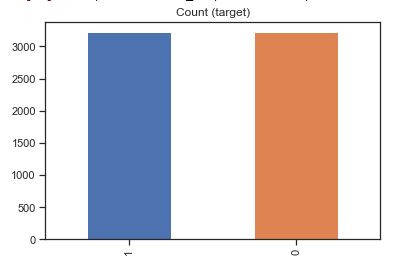


4. Oversampling

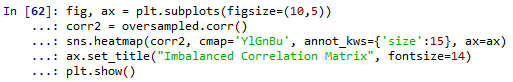
Sekarang akan kita perlakukan dataset kita sebagai imbalanced data. Perlu dilakukan resampling agar training data menjadi balanced dan mengurangi risiko hasil yang bias pada kelas mayoritas. Dipilih metode oversampling dikarenakan jumlah data yang masih dibawah puluhan ribu, sehingga pengolahan data pun masih memungkinkan dan relatif cepat.

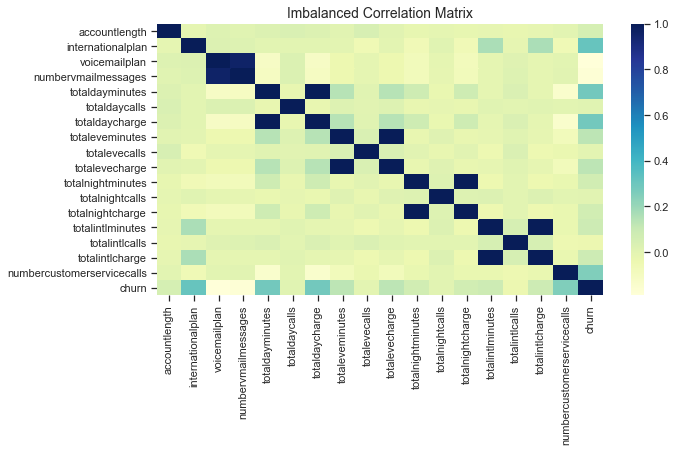
Oversampling dapat didefinisikan sebagai menambahkan lebih banyak salinan dari kelas minoritas. Oversampling bisa menjadi pilihan yang baik ketika kita tidak memiliki banyak data untuk dikerjakan. Akan digunakan modul resampling dari Scikit-Learn untuk mereplikasi sampel secara acak dari kelas minoritas. Catatan penting: selalu bagi menjadi data test set dan train set SEBELUM mencoba teknik oversampling! Oversampling sebelum pemisahan data dapat memungkinkan pengamatan yang sama persis untuk muncul di kedua set. Hal ini menyebabkan overfitting dan generalisasi yang buruk pada test data.





Setelah melakukan resampling, kita memiliki rasio poin data yang sama untuk setiap kelas. Akan kita lihat korelasi antar variabelnya.





Perhatikan bahwa korelasi fitur jauh lebih jelas sekarang. Sebelum memperbaiki masalah ketidakseimbangan, sebagian besar fitur tidak menunjukkan korelasi yang pasti akan berdampak pada kinerja model. Karena korelasi fitur sangat penting untuk kinerja model secara keseluruhan, maka dari itu penting untuk memperbaiki ketidakseimbangan karena juga akan berdampak pada kinerja model.

# Analysis and Interpretation

1. Metrics

Sekarang kita akan menganalisa dan memodelkan data train dari langkah processing tadi, model classifier mana yang paling sesuai untuk memprediksi churn pelanggan. Hal ini dapat dilakukan dengan pemilihan metrics yang paling sesuai.

Salah satu masalah utama ketika berhadapan dengan imbalanced dataset berkaitan dengan metrik yang digunakan untuk mengevaluasi model kita. Menggunakan metrik yang lebih sederhana seperti akurasi dapat menyesatkan. Dalam dataset dengan kelas yang sangat tidak seimbang, jika model classifier selalu "memprediksi" kelas yang paling umum tanpa melakukan analisis fitur, itu masih akan memiliki tingkat akurasi yang tinggi.

Cara yang menarik untuk mengevaluasi model classifier adalah dengan menggunakan confussion matrix, yang menunjukkan prediksi yang benar dan salah untuk setiap kelas.

* Di baris pertama, kolom pertama menunjukkan berapa banyak kelas 0 yang diprediksi dengan benar: True Negative
* Di baris pertama, kolom kedua menunjukkan berapa banyak kelas 0 yang diprediksi sebagai 1: False Positive.
* Di baris kedua, kolom pertama mencatat banyak kelas 1 diprediksi secara keliru sebagai kelas 0: False Negative.
* Di baris kedua, kolom kedua mencatat banyak kelas 1 diprediksi secara benar: True Positive.

Oleh karena itu, semakin tinggi nilai diagonal dari confusion matrix semakin baik, menunjukkan banyak prediksi yang benar.

Dalam kasus ini, kita tidak menginginkan False Negative terjadi karena kerugian perusahaan lebih besar jika pelanggan churn diprediksi tidak churn🡪 kehilangan pelanggan. Kita akan lebih menerima kesalahan False Positive.

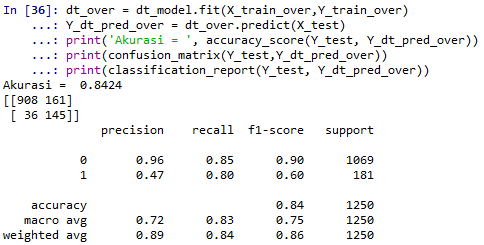
Sehingga dipilih model yang metrics Recall tinggi dibandingkan akurasi tinggi, di mana recall merupakan rasio prediksi benar positif dibandingkan dengan keseluruhan data yang benar positif.

Akan lebih baik juga apabila F1 Score tinggi, di mana F1 scoremerupakan perbandingan rata-rata presisi dan recall yang dibobotkan F1 Score = 2 x (Recall x Precission) / (Recall + Precission)

2. Model Classifier

Sekarang kita coba modelkan DT, RF, NB dengan training data yang sudah seimbang.

* Decision Tree

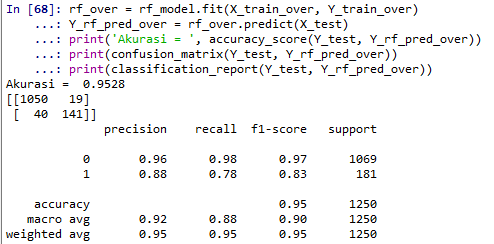


Akurasi model cukup tinggi, yaitu 84,24%; %; ini disebabkan hasil prediksi benar bias terhadap lebel mayoritas yaitu True Negative besar = 908. Walaupun masih relatif besar namun perhatikan bahwa angka ini lebih baik dibandingkan model dengan training data yang imbalanced. Bisa kita lihat F1 scorenya saja yaitu 0,6.

Presisi: terdapat 47% pelanggan yang benar churn dari keseluruhan pelanggan yang diprediksi churn 🡪 63% pelanggan yang sebenarnya tidak churn akan diperlakukan sebagai churn. Jika perusahaan akan memberikan perlakuan khusus untuk pelanggan yang diprediksi churn, seperti pemberian promo, hal ini jadi pertimbangan untuk tidak memilih model ini. Angka kesalahan yang cukup besar dan tidak tepat sasaran ini juga dapat merugikan perusahaan.

Dengan recall dari churn yang lebih baik dari DT model yaitu 0,8; hal ini sangat baik karena true churn lebih banyak diprediksi sebagai churn yaitu 145 dari 181 pelanggan. Namun dikarenakan presisi yang rendah, model ini tidak dianjurkan untuk digunakan.

* Random Forest

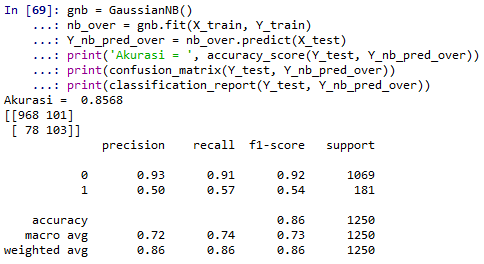


Akurasi model tinggi, yaitu 95,28%; hal ini juga dikarenakan True Negative besar = 1050. Bisa kita lihat F1 scorenya saja yaitu 0,83 yang bisa dibilang cukup baik.

Presisi: terdapat 88% pelanggan yang benar churn dari keseluruhan pelanggan yang diprediksi churn, angka ini sebenarnya sudah cukup baik 🡪 12% pelanggan yang sebenarnya tidak churn akan diperlakukan sebagai churn. Jika dilihat dari presisi, model ini lebih baik daripada DT karena lebih sedikit yang “salah target”.

Dengan recall yaitu 0,78; hal ini cukup baik karena true churn lebih banyak diprediksi sebagai churn yaitu 141 dari 181 pelanggan. Sehingga model ini sudah baik untuk memprediksi churn.

* Naïve Bayesian



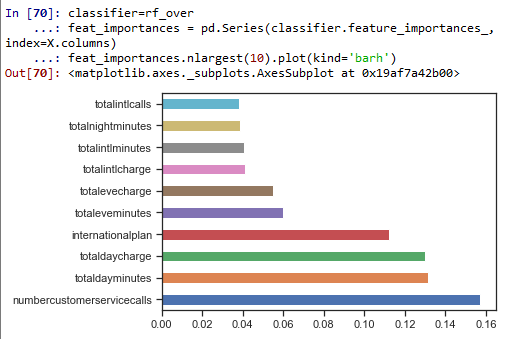
Akurasi model cukup tinggi, yaitu 85,68%; hal ini juga dikarenakan True Negative besar = 968 dibandingkan dengan True Positive. Bisa kita lihat F1 scorenya saja yaitu 0,54.

Presisi: terdapat 50% pelanggan yang benar churn dari keseluruhan pelanggan yang diprediksi churn, angka ini sebenarnya sudah cukup baik 🡪 50% pelanggan yang sebenarnya tidak churn akan diperlakukan sebagai churn. Jika dilihat dari presisi, model ini tidak dapat dikatakan baik. Dengan recall yaitu 0,57; angka ini rendah jika dibandingkan dengan kedua model sebelumnya.

Dari hasil interpretasi di atas, kita telah dapatkan model yang paling baik dari ketiga model yang kita coba kali ini adalah Random Forest.

3. Variable Importances

Selanjutnya ingin diketahui variabel prediktor mana yang paling memengaruhi churn pelanggan. Hal ini dapat dilihat dari plot 10 variabel yang paling berpengaruh berikut.



Jumlah panggilan ke customer service, jumlah menit panggilan saat siang hari, besar biaya yang ditagih oleh Telecom ke pelanggan pada siang hari, dan kepunyaan-atau-tidaknya international plan pelanggan ternyata adalah variabel-variabel yang paling memengaruhi dibanding variabel lainnya.

# Conclusion

Untuk menyimpulkan hasil analisis, kita kembali ke tujuan kita adalah yaitu mengidentifikasi pelanggan yang cenderung churn. Sehingga, kita dapat menganjurkan Telecom untuk melakukan strategi pemasaran yang bertujuan khusus pada menghindari peristiwa churn. Untuk itu, kita telah mengevaluasi data dan melakukan pengklasifikasian yang berbeda. Kita telah menerapkan teknik oversampling untuk menangani variabel target yang tidak seimbang. Dalam tahapan pengklasifikasi kita juga telah melakukan beberapa pengklasifikasi yang berbeda, termasuk Decision Tree, Random Forest, dan Naïve Bayes. Ditemukan bahwa kinerja terbaik berdasarkan F1 score, presisi, dan recall dicapai oleh Random Forest.

Saran untuk Telecom:

Prediksi churn pelanggan sangat penting untuk stabilitas keuangan jangka panjang perusahaan. Dari hasil analisis, salah satu prediktor paling penting untuk Random Forest adalah jumlah panggilan ke customer service. Ini mungkin menyiratkan bahwa perusahaan harus meningkatkan layanan pelanggannya. Fitur penting lainnya adalah jumlah menit panggilan saat siang hari, besar biaya yang ditagih oleh Telecom ke pelanggan pada siang hari. Jadi perusahaan dapat mencoba menurunkan biaya per menit untuk klien, yang memiliki beberapa menit sehari atau dapat menawarkan tarif tetap untuk panggilan, khusus wakti siang hari.

Saran untuk Analis lain:

Metode oversampling pada laporan ini hanya beberapa dari banyak metode yang mungkin untuk dicoba ketika berhadapan dengan set data yang tidak seimbang. Beberapa metode lain yang perlu dipertimbangkan adalah undersampling, mengumpulkan lebih banyak data atau memilih rasio resampling yang berbeda—tidak harus memiliki rasio 1: 1! Begitu pula dengan metode klasifikasi, masih banyak pilihan model classifier lain yang bisa dicoba. Analis selanjutnya harus selalu mencoba beberapa pendekatan dan kemudian memutuskan mana yang terbaik untuk tiap masalah tertentu.

# Reference

<https://www.kaggle.com/ambpro/dealing-with-unbalance-eda-pca-smote-lr-svm-dt-rf>

<https://www.kaggle.com/rafjaa/resampling-strategies-for-imbalanced-datasets>

<https://towardsdatascience.com/methods-for-dealing-with-imbalanced-data-5b761be45a18>

<https://academy.vertabelo.com/blog/python-customer-churn-prediction/>

<https://towardsdatascience.com/accuracy-recall-precision-f-score-specificity-which-to-optimize-on-867d3f11124>