Introduction

Jika Anda memiliki rekening bank, mungkin Anda sudah tahu bahwa bank-bank menawarkan bonus pembukaan berupa cashback atau poin hadiah untuk mendorong calon klien. Bonus pembukaan yang umumnya berjalan dengan cara bahwa jika klien setuju untuk menyimpan sejumlah aset tertentu di bank, ia akan menerima potongan langsung atau promosi untuk layanan-layanan lain dari bank. Jika Anda adalah klien bank berpengalaman, mungkin Anda pernah dihubungi oleh bank-bank mengenai penawaran retensi. Bank-bank memberikan berbagai penawaran keuangan kepada klien untuk mencegah mereka pergi. Ada dua alasan utama mengapa bank begitu obsesif dengan penawaran-penawaran ini. Pertama, bank perlu menjaga hubungan yang sehat dengan klien untuk menjaga peningkatan pendapatan yang konsisten. Kedua, industri keuangan sangat kompetitif. Bank-bank terus meluncurkan program pemasaran baru untuk mencegah klien-klien pergi ke pesaing-pesaing. Setelah beberapa dekade peningkatan, saat ini kebanyakan bank komersial telah mengadopsi retention-based client relationship model.

Proyek ini menerapkan teknik-teknik pembelajaran mesin berbimbing untuk mengembangkan model prediksi yang proaktif untuk memprediksi apakah klien bank komersial akan tetap tinggal atau tidak. Data berasal dari catatan transaksi nyata dari Xiamen International Bank, sebuah bank komersial utama di China. Proyek ini dimulai dengan Exploratory Data Analysis (EDA), diikuti oleh model development and implementation. Proyek ini kemudian memperkirakan retention rate klien menggunakan model tersebut. Akhirnya, proyek ini menyediakan solusi bisnis mengenai bagaimana menargetkan klien dengan tepat dan efisien yang cenderung Churn/berpindah.

Business Understanding

Ketika kita memasuki data-driven era, bank-bank komersial dihadapkan pada persaingan yang semakin meningkat di seluruh dunia. Pandemi pada tahun 2020 menambah beban berat tambahan pada sistem hubungan klien bank yang sudah kewalahan. Untuk menjaga peningkatan pendapatan, bank-bank memerlukan pemahaman dan estimasi yang lebih baik mengenai permintaan dan preferensi klien. Secara khusus, bank tertarik untuk memprediksi tingkat perpindahan klien dan perubahan minat keuangan mereka. Melalui penargetan dan pemasaran, bank dapat mengurangi kerugian pendapatan dengan mempertahankan klien yang hendak pergi.

Untuk secara efektif mempertahankan klien, bank-bank telah mendirikan berbagai model bisnis berbasis retensi untuk menjaga loyalitas klien. Model-model ini dimulai dengan meluncurkan insentif-insentif awal yang beragam untuk menarik calon klien potensial. Bank-bank menggunakan teknik pemasaran untuk menargetkan dengan tepat klien-klien yang tertarik pada produk mereka. Setelah bank mendapatkan klien, bank mulai mengumpulkan sebanyak mungkin informasi tentang klien tersebut. Kategori informasi meliputi aset dan perilaku hingga informasi pribadi. Kemudian, bank memprediksi apakah klien akan tetap tinggal, berdasarkan informasi yang telah terkumpul. Dengan kata lain, bank perlu memprediksi apakah seorang klien akan pergi dan beralih ke pesaing, atau "churn". Perpindahan ini tidak hanya secara langsung memotong sumber pendapatan, tetapi juga mengurangi minat klien potensial. Jika seorang klien memutuskan untuk pergi, kemungkinan besar ia akan meyakinkan orang-orang yang ia kenal

untuk tidak memilih produk atau layanan dari bank tersebut. Bank-bank mengalami kerugian pendapatan yang berlipat ganda akibat satu aktivitas churn. Oleh karena itu, untuk mencegah hal ini, bank-bank bersedia mengorbankan sedikit laba dengan menawarkan insentif retensi kepada klien-klien yang mungkin akan pergi/Churn.

Pada pandangan pertama, model ini efektif: bank meningkatkan anggaran pemasaran untuk mempertahankan klien yang kemungkinan akan pergi, tetapi akhirnya menghasilkan keuntungan yang cukup untuk menutupi biaya tambahan. Namun, dengan pandangan yang lebih mendalam, model retensi menimbulkan masalah-masalah baru. Pertama, tidak semua klien yang berisiko akan pergi. Beberapa klien mungkin hanya mencoba-coba dan melihat apakah mereka dapat mendapatkan manfaat tambahan. Beberapa mungkin akan tetap tinggal selama setahun lagi meskipun pengukuran mereka menunjukkan bahwa mereka sangat mungkin pergi. Kedua, beberapa klien akan pergi tanpa peduli dengan penawaran keuangan apa pun yang mereka terima. Misalnya, jika seorang klien memutuskan untuk selamanya meninggalkan negara, maka ia kemungkinan akan menutup rekening, tanpa memedulikan penawaran retensi apa pun. Jika seorang klien memiliki sejumlah besar uang dan memutuskan untuk pergi, ia akan kurang tertarik pada insentif keuangan. Efek dari penawaran retensi kemudian akan menjadi sangat kecil.

Penting bagi bank-bank untuk merancang dan mengimplementasikan mekanisme prediksi yang secara aktif memperkirakan risiko perpindahan klien dan melakukan intervensi proaktif sebelum klien tersebut membuat keputusan akhir. Ini adalah tujuan utama dari proyek ini: menerapkan teknik advanced machine learning untuk membangun model prediksi.

Research Questions

1. Apa fitur-fitur paling penting dari data? Apakah ada temuan menarik mengenai data?
2. Algoritma mana yang memiliki kinerja terbaik dalam memprediksi perpindahan? Bagaimana cara mengestimasikan kinerja tersebut
3. Bagaimana cara menerapkan hasil model untuk memecahkan masalah perpindahan di dunia nyata?

Data and Sample

(https://www.kaggle.com/datasets/shangweichen/xiamen-international-bank-modeling-competition)

Data berasal dari Xiamen International Bank, sebuah bank komersial utama di China. Sampel yang tersedia berisi daily transactions records, dalam berbagai kategori, untuk kuartal ketiga dan keempat tahun 2019. Terdapat tiga set data:

● Train set:
x_train : this is the train set. It contains all available features (predictors)
● Test set:
x_test : this is the test set. It contains the same features as x_train.
● Validation set:
Y_train : this is the validation set. It contains the results (also known as

the label) of the x_train . In this case, the results are indicators of whether or not clients churn.

The train set dan the validation set diambil secara acak dari transaction records kuartal ketiga dan keempat tahun 2019. The test set diambil secara acak dari kuartal pertama tahun 2020.

Data ini mengandung 55 features. Ada lima kategori features dalam train set dan test set:

1. “X” (8 features): this category includes information regarding client’s assets at the end of each month. Features include structured deposit balance, loan balance, financial products balance, and so on. [L SEP]

Variable Name	Description
cust_no	custumer's ID (primary key)
X1	structured deposit balance
X2	time deposit balance
X3	demand deposit balance
X4	financial products balance
X5	fund balance
X6	asset management balance
X7	loan balance
X8	large deposit certificate balance

2. “B” (7 features): this category records client’s behaviors in each month. Features include number of transfers, latest transfer date/time, transfer amounts, and so on.

Variable Name	Description
cust_no	custumer's ID (primary key)
B1	mobile banking login times
B2	transfer-in times
B3	transfer-in money amount
B4	transfer-out times
B5	transfer-out money amount
B6	latest transfer time
B7	number of transfers in a season

3. “E” (18 features): this category records client’s important behaviors in each season, such as first time loan date/time, first overdue date, first online banking login date, and so on.

Variable Name	Description
cust_no	custumer's ID (primary key)
E1	account opening date
E2	online banking opening date
E3	mobile banking opening date
E4	first online banking login date
E5	first mobile banking login date
E6	first demand deposit date
E7	first time deposit date
E8	first loan date
E9	first overdue date
E10	first cash transaction date
E11	first bank-securities transfer date
E12	first transfer at counter date
E13	first transfer via online banking date
E14	first transfer via mobile banking date
E15	maximum amount transferred out of another bank
E16	maximum amount transferred out of another bank date
E17	Maximum transfer amount from other bank
E18	Maximum transfer amount from other bank date

4. “Y” (2 features): this category contains client’s deposits in each month.

Variable Name	Description
cust_no	custumer's ID (primary key)
C1	deposit products value
C2	number of deposit products

5. “I” (20 features): this category contains client’s information (trivias) in each season.  Features include gender, age, occupation, education level, and so on.

l1	gender
l2	age
l3	class
l4	tag
l5	occupation
l6	deposit customer tag
l7	number of products owning
l8	constellation
l9	contribution
l10	education level
l11	family annual income
l12	field description
l13	marriage description
l14	occupation description
l15	QR code recipient
l16	VIP
l17	online banking client
l18	mobile banking client
l19	SMS client
l20	WeChat Pay client

The test set has two columns:
1. “Cust_no”: customer’s unique ID
2. “label”: whether or not a customer churns. There are three possible values:

- 1: indicates churn.
- -1: indicates not churn
- 0: indicates no preference. [SEP]

Dimensions

The raw train set contains 465,441 rows and 56 columns (1 index and 55 features). The train validation set contains 145,296 rows and 56 columns. The test set contains 76,722 rows and 1 index column.

Model

1. Random Forest
2. Logistic Regression with Elastic Net
3. Gradient Boosting Machin