Bambang Sugeng Marsudianto

Accelarated Machine Learning

Topic 1 2 Assignment | Data Science Use Cases and Creating Github Account

Sumber : Keldine, M.,2018 “Bank Customer Churn Prediction”., Kaggle
https://www.kaggle.com/code/kmalit/bank-customer-churn-prediction

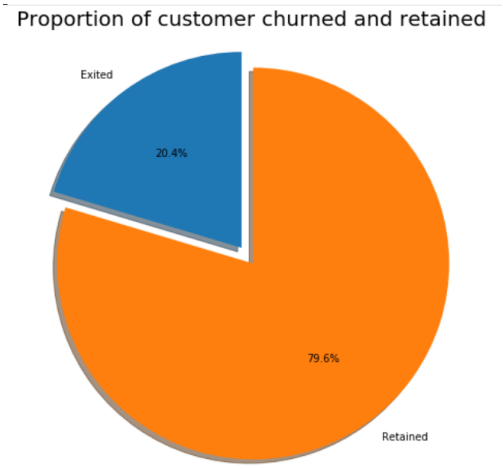
Pendahuluan :

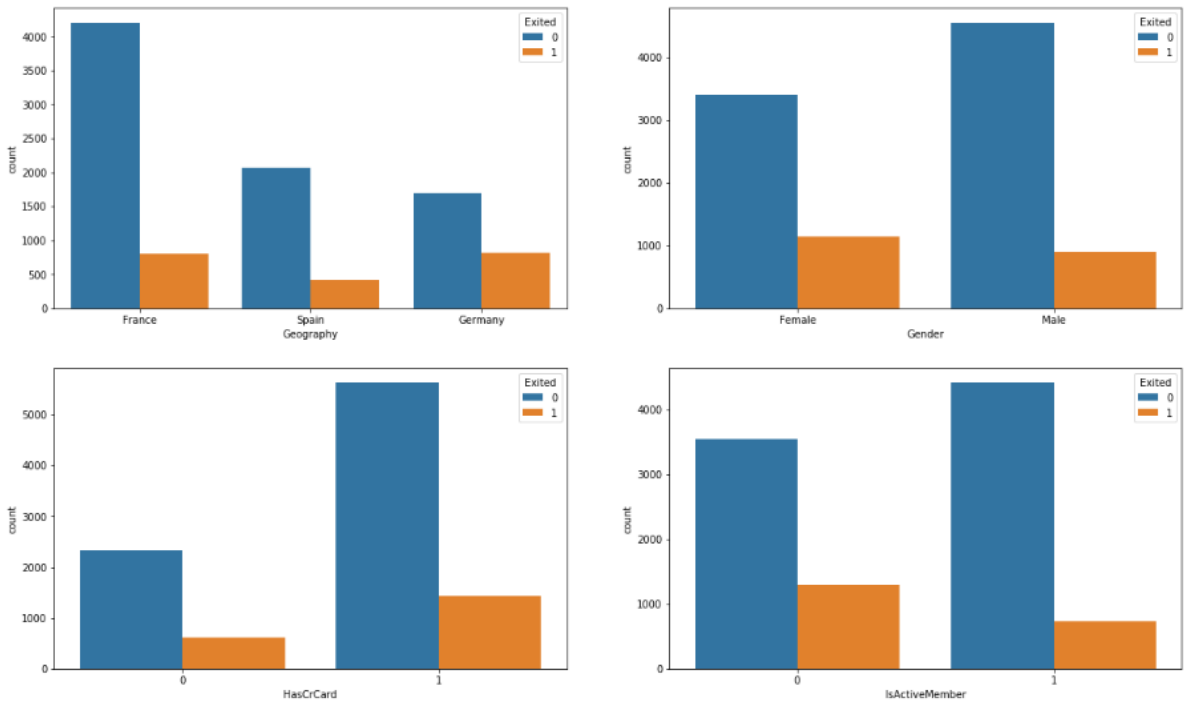
Pada sumber tersebut penulis ingin mengidentifikasi dan mevisualisasikan apa saja factor yang mengakibatkan bank mengalami churn terhadap para pelanggan atau nasabah. Selain itu beliau juga membuat pemodelan pemodelan yang berguna untuk memprediksi. Pemodelan yang akan digunakan yaitu mengklasifikasi pelanggan akan melakukan churn atau tidak, probabilitas model yang akan melakukan churn berguna untuk memudahkan layanan pelanggan.

Isi :

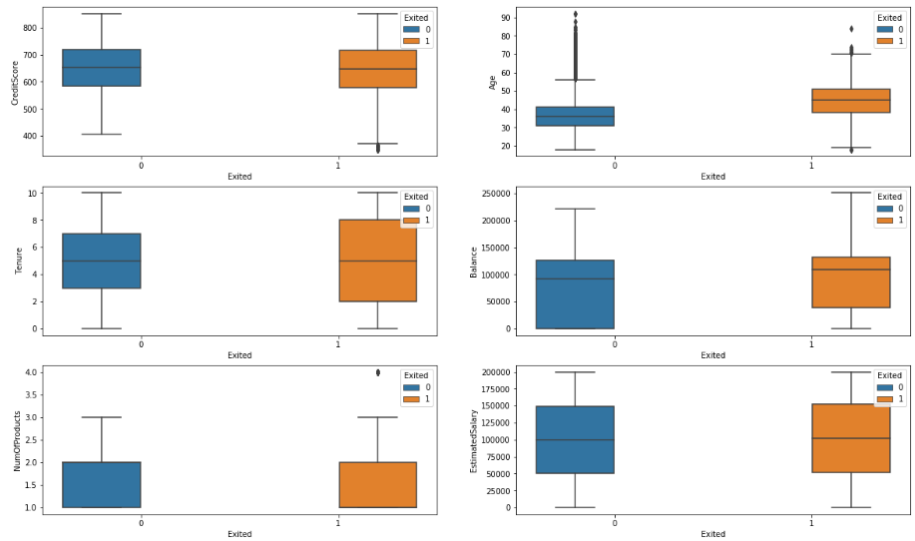
Pertama – tama data yang sudah ada direview dan di siapkan. Mereka mengimport library dan juga mengambil data atau mengate data yang sudah disiap kan. Data berjumlah 1000 baris dan 14 atribut. Dari data yang sudah diketahui penulis tidak memerlukan 2 atribut, dan juga nama keluarga karena ini akan menghasikan profil jadi mengecualikan itu juga. Ketika memunculkan data 5 teratas baru penulis muncul pertanyaan – pertanyaan, contoh seperti Ada pelanggan yang sudah exit tapi masih ada saldo di akunnya! Maksudnya gimana ini?.

Pada bagian kedua penulis mengeksplorasi data dan menganalisis yang ada atau biasa disebut EDA (Exploratory Data Analysis). Pada bagian ini beliau berfokus bagaimana atribut itu ketika status nya “keluar” ditampilkan dalam table pie jumlah keluar sebanyak 20%. Angka 20% angka yang kecil beliau perlu mamastikan bahwa model yang dipilih benar – benar akurat karena menarik bagi bank untuk mengidentifikasi dan mempertahankan kelompok churns sebagai lawan prediksi pelanggan yang harus dipertahankan



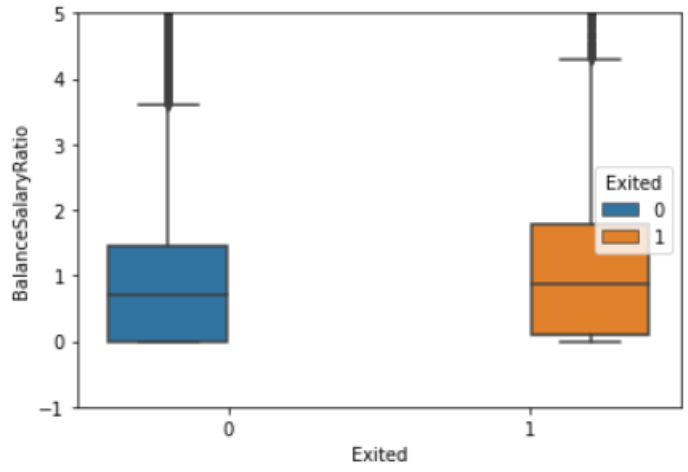


Pada grafik ini bisa kita liar pelanggan kebanyakan dari francis. Namun, proposi pelanggan yang churn berbanding terbalik dengan populasi pelanggan yang mengacu pada bank yang mungkin memiliki masalah, masalah nya mungkin tidak cukup besar layanan yang diberikan diarea dimana mereka ia memiliki lebih sedikit klien. Pelanggan wanita lebih besar mengalami churn dibandingkan pelanggan wanita. Mereka yang mengami churn adalah mayoritas yang mempunyai kartu kredit. Mereka yang tidak aktif juga mengalami churn lebih besar.karena yang tidak aktif lumayan banyak mungkin bank harus merayu agak para pelanggan untuk mengaktifkan nya kembali karena bisa berdampak positif pada churn nasabah.



Pada grafik diatas ada beberapa point yang disampaikan. Tidak ada perbedaan yang signifikan dalam distribusi skor kredit antara pelanggan yang dipertahankan dan dihentikan. Pelanggan yang lebih tua lebih banyak yang mengalami churn dibandingkan pelanggan yang muda.

Selanjutnya adalah feature engineering yaitu untuk menambahkan fitur yang mungkin yang berdampak pada kemungkinan churning. Seperti biasa membagi antara data tes dan data train



Pada grafik ini dijelaskan bahwa gaji memiliki pengaruh kecil terhadap kemungkinan pelanggan churning, namun rasio saldo bank dan perkiraan gaji menunjukkan bahwa pelanggan dengan rasio gaji saldo yang lebih tinggi lebih banyak yang akan mengkhawatirkan bank karena hal ini berdampak pada sumber modal pinjaman mereka.

Membuat variable baru karena usia mempengaruhi gaji yang bertujuan untuk menstandarkan masa kerja di atas usia.

```
# Resulting Data Frame
df_train.head()
```

NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited	BalanceSalaryRatio	TenureByAge	CreditScoreGivenAge
2	1	1	15306.29	0	0.000000	0.240000	18.440000
1	1	1	20555.21	0	4.398550	0.114286	17.685714
1	1	0	102085.35	0	1.195454	0.200000	17.475000
1	1	1	111184.67	0	1.117305	0.048780	13.609756
1	0	1	155643.04	0	0.857741	0.147059	18.764706

Selanjutnya data prep for model fitting. Memilih atribut untuk membuat data training dan untuk data training isinya gambar dibawah ini.

```
# Arrange columns by data type for easier manipulation
continuous_vars = ['CreditScore',  'Age', 'Tenure', 'Balance','NumOfProducts', 'EstimatedSalar
y', 'BalanceSalaryRatio',
                   'TenureByAge','CreditScoreGivenAge']
cat_vars = ['HasCrCard', 'IsActiveMember','Geography', 'Gender']
df_train = df_train[['Exited'] + continuous_vars + cat_vars]
df_train.head()
```

	Exited	CreditScore	Age	Tenure	Balance	NumOfProducts	EstimatedSalary	BalanceSalaryRatio	TenureByAge
8159	0	461	25	6	0.00	2	15306.29	0.000000	0.240000
6332	0	619	35	4	90413.12	1	20555.21	4.398550	0.114286

Mengubah 0 menjadi -1 sehingga model dapat menangkap hubungan yang benilai negatf dimana atribut tidak diterapkan dan juga mengganti nama negara, jenis kelamin jika bernilai true maka akan 1 dan -1 bernilai false. Dilanjutkan dengan minmax scaller dan prep pipeline for test data.

HasCrCard	IsActiveMember	Geography_Spain	Geography_France	Geography_Germany	Gender_Female	Gender_Male
1	1	1	-1	-1	1	-1
1	1	-1	1	-1	1	-1
1	-1	-1	1	-1	1	-1
1	1	-1	-1	1	-1	1
-1	1	-1	1	-1	-1	1

Pada bagian terakhir yaitu model fitting and selection. Disini akan dilakukan logistic regeresion, SVM dan Ensembel models.

```
# Fit primal logistic regression
param_grid = {'C': [0.1,0.5,1,10,50,100], 'max_iter': [250], 'fit_intercept':[True],'intercept_s
caling':[1],
              'penalty':['l2'], 'tol':[0.00001,0.0001,0.000001]}
log_primal_Grid = GridSearchCV(LogisticRegression(solver='lbfgs'),param_grid, cv=10, refit=True,
verbose=0)
log_primal_Grid.fit(df_train.loc[:, df_train.columns != 'Exited'],df_train.Exited)
best_model(log_primal_Grid)
```

```
# Fit SVM with RBF Kernel
param_grid = {'C': [0.5,100,150], 'gamma': [0.1,0.01,0.001],'probability':[True],'kernel': ['rb
f']}
SVM_grid = GridSearchCV(SVC(), param_grid, cv=3, refit=True, verbose=0)
SVM_grid.fit(df_train.loc[:, df_train.columns != 'Exited'],df_train.Exited)
best_model(SVM_grid)
```

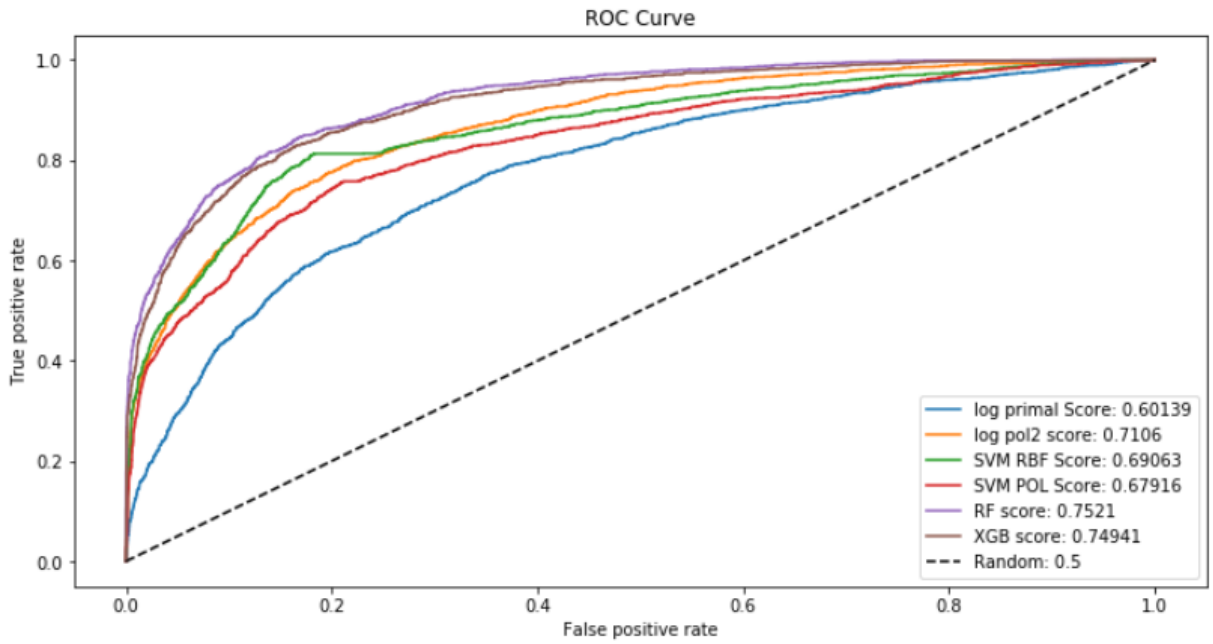
Ulasan akurasi kecocokan model terbaik : Ketertarikan pada kinerja dalam memprediksi 1 (Pelanggan yang churn) sebagai contoh berikut dibawah ini.

```
print(classification_report(df_train.Exited, log_primal.predict(df_train.loc[:, df_train.columns
!= 'Exited'])))
```

```
              precision    recall  f1-score   support

           0       0.83      0.97      0.89      6353
           1       0.64      0.24      0.35      1647

   micro avg       0.82      0.82      0.82      8000
   macro avg       0.73      0.60      0.62      8000
weighted avg       0.79      0.82      0.78      8000
```



Dari table diatas tujuan utama penulis adalah memprediksi pelanggan yang mungkin akan churn. Dari review model-model yang dipasang di atas, model terbaik yang memberikan keseimbangan yang layak antara recall dan precision adalah random forest dimana menurut fit pada training set, dengan nilai presisi pada 1 sebesar 0,88, dari semua pelanggan yang model berpikir akan churn, 88% benar-benar churn dan dengan skor recall 0,53 pada 1, model mampu menyoroti 53% dari semua yang churn.

Kesimpulan :

Kita dapat mengetahui step by step cara memprediksi churn yang ada dibank. Untuk memudahkan mencari atau mengambil data apa saja data yang diperlukan agar lebih muda kita bisa gunakan grafik. Kita jadi lebih mengetahui cara modeling fitting yang akan digunakan. Dari uraian yang saya buat juga saya tau sebanyak model yang memiki akurasi tinggi masih bisa meleset sekitar setengah dari mereka yang churning. Ini bisa dilakukan pelatihan ulang model dengan lebih banyak data.

Github repository

https://github.com/SugengMars/data-science-testing-zenius-task