VIX DATA SCIENTIST
HOME CREDIT X RAKAMIN ACADEMY

DISUSUN OLEH: GISELLE HALIM

PROBLEM RESEARCH

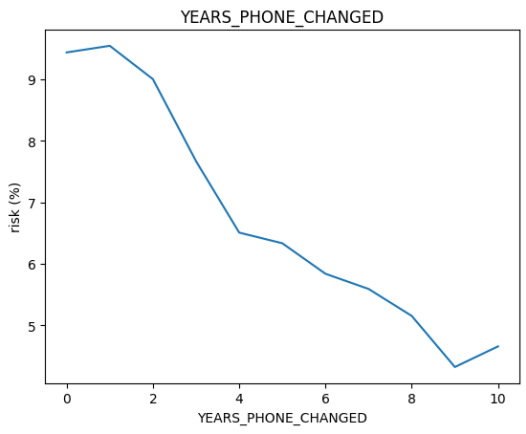
Home credit membutuhkan business analysis dan model machine learning yang dapat memberikan insight untuk menerima ajuan dari peminjam yang berpotensi baik melunaskan pinjamannya. Untuk mencapai tujuan ini, dibutuhkan analisis mengenai pattern yang dimiliki peminjam yang baik serta bantuan dari model machine learning untuk memproses data yang banyak. Dapat juga dilakukan penyaringan untuk mencegah peminjam yang berpotensi tidak melunaskan pinjaman.

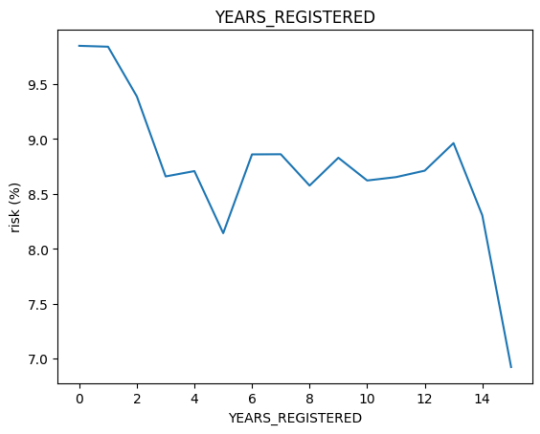
Selain itu, dibutuhkan analisa terhadap demografis peminjam untuk mengetahui target customer yang berpotensi untuk mengajukan pinjaman kedepannya.

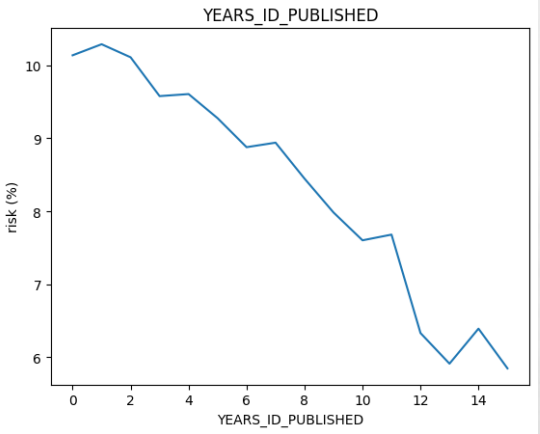
DATA PRE-PROCESSING

- Menghapus column yang tidak diperlukan dalam proses analysis.
- Membuat column baru berdasarkan column yang menghitung hari. Ini dilakukan untuk mengetahui usia, lama bekerja, waktu sejak penggantian nomor telepon, dan lainnya.
- Membuat column baru untuk menghitung berapa banyak dokumen yang di-flag agar tidak terlalu banyak column pada dataset.
- Melakukan imputasi untuk mengisi data yang kosong/null.
- Melakukan label encoder untuk mengubah data kategorikal menjadi numerik agar dapat digunakan dalam machine learning.

DATA VISUALIZATION AND BUSINESS INSIGHT

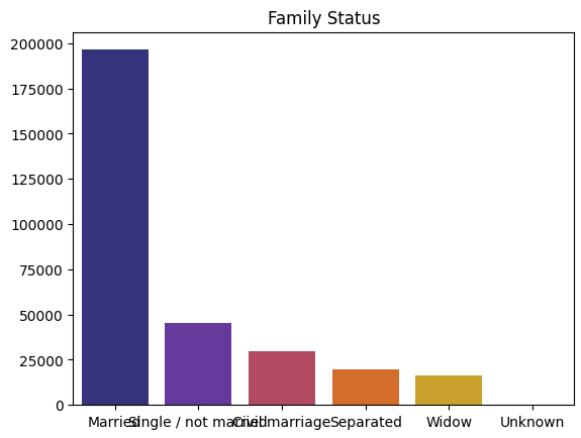






Jangka waktu sejak pergantian dokumen, identitas, dan nomor telepon terakhir dapat menentukan resiko peminjaman. Semakin singkat jarak pergantiannya (banyak mengubah dokumen dan data diri), maka resiko akan meningkat. Dapat diadakan screening tambahan untuk memastikan pergantian tersebut bukanlah untuk data palsu, dsb.

DATA VISUALIZATION AND BUSINESS INSIGHT



	CODE_GENDER	NAME_FAMILY_STATUS	Count	Percent
3	F	Married	122445	0.398181
4	M	Married	73984	0.240590
9	F	Single / not married	28584	0.092953
0	F	Civil marriage	20769	0.067539
10	M	Single / not married	16860	0.054827

Peminjam mayoritas telah berstatus menikah, dapat dijadikan target promosi karena pengeluaran meningkat (terutama jika sudah memiliki anak). Berkaitan dengan ini, peminjam yang berstatus menikah lebih banyak perempuan yang biasanya mengurus keperluan rumah tangga.

DATA VISUALIZATION AND BUSINESS INSIGHT

Insight Lainnya:

- Rata-rata peminjam membayar dengan tepat waktu.
- Peminjam mayoritas jenis kelamin perempuan.
- Mayoritas peminjam adalah laborers/buruh yang pemasukannya lebih rendah, sehingga bisa dijadikan target promosi untuk melakukan pinjaman.
- Semakin besar rating region client (berdasarkan standar yang telah ditentukan sebelumnya), resiko peminjaman tidak baik semakin tinggi.
- Peminjam yang belum lama bekerja memiliki resiko lebih tinggi dari yang sudah bekerja lama.
- Peminjam yang memiliki 3 sampai 4 dokumen yang di-flagged, berpotensi jauh lebih besar untuk melakukan pinjaman dengan tidak baik.
- Hampir sebagian peminjam memiliki rumah dan tidak memiliki mobil.

MACHINE LEARNING IMPLEMENTATION

```
[331] #Defining x and y
      x = df.drop(columns=['TARGET'], axis = 1)
      y = df['TARGET']
```

```
[332] from sklearn.model_selection import train_test_split
      x_train, x_test, y_train, y_test = train_test_split(x, y, test_size = 0.2, random_state = 42)
```

```
[333] x_train.shape, x_test.shape

      ((246008, 42), (61503, 42))
```

Balancing data

The data is unbalanced, so we need to balance it first by oversampling the minority.

```
[334] from imblearn.over_sampling import SMOTE

      SMOTE = SMOTE()
      x_train, y_train = SMOTE.fit_resample(x_train, y_train)
```

```
[335] from sklearn.ensemble import RandomForestClassifier
      rf = RandomForestClassifier(max_depth=5)
```

Split data training dan testing dilakukan dengan rasio 80:20. Karena data imbalance, dilakukan oversampling dengan SMOTE untuk menyeimbangkan kelas minoritas.

Model di-training dengan algoritma random forest.

MACHINE LEARNING EVALUATION

```
#Check model performance using classification_report
print(classification_report(y_test, y_pred))
```

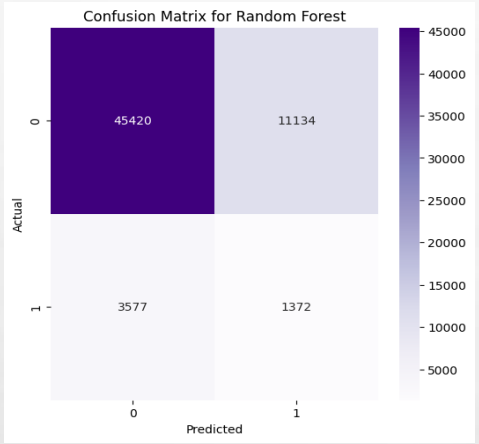
```
              precision    recall  f1-score   support

           0       0.93      0.80      0.86     56554
           1       0.11      0.28      0.16      4949

    accuracy                           0.76     61503
   macro avg       0.52      0.54      0.51     61503
weighted avg       0.86      0.76      0.80     61503
```

```
#Check model performance using auc score
roc_auc_score(y_test, y_pred)*100
```

```
54.017696921228705
```




```
print('Training-set accuracy score:', rf.score(x_train, y_train))
print('Test-set accuracy score:', rf.score(x_test, y_test))
```

```
Training-set accuracy score: 0.8221238037960129
Test-set accuracy score: 0.7608084158496333
```

Hasil evaluasi dari model random forest tidak terlalu baik, dapat terlihat model kesulitan untuk memprediksi kelas ‘1’ atau peminjam yang tidak baik. Hal ini dapat dikarenakan oleh ketidakseimbangan data, sehingga membutuhkan perbaikan lebih lanjut.

LINK GITHUB

TERIMA KASIH