Final Project

Bootcamp SVI

Daffa Abiyyu

Business Understanding

Business Objectives

Perusahaan ingin menawarkan produk 7 kepada pelanggan, produk 7 adalah produk vending machine yang ditawarkan pada toko-toko retail.

Model Objectives

Membuat mesin klasifikasi untuk menentukan apakah seorang pelanggan sebaiknya ditawarkan **produk 7** atau tidak.

Model Success Criteria

Recall minimal 0,7 dan FPR maksimal 0,3.

Data Understanding

Data Description

- umur : umur pelanggan
- kota : kota tempat tinggal pelanggan
- keluarga : status keluarga pelanggan
- Pekerjaan : pekerjaan pelanggan
- gender : jenis kelamin pelanggan
- punya_produk : apakah pelanggan memiliki produk atau tidak
- average_sisauang_mingguan : rata-rata sisa uang mingguan pelanggan
- sisauang_akhir : sisa uang akhir pelanggan
- sisauang_tertahan : sisa uang tertahan pelanggan
- sisauang_tersedia : sisa uang tersedia pelanggan
- sisa_hutang : sisa hutang pelanggan
- jumlah_pembayaran_hutang : jumlah pembayaran hutang pelanggan
- jumlah_pengeluaran : jumlah pengeluaran pelanggan
- jumlah_pemasukan : jumlah pemasukan pelanggan
- frekuensi_pengeluaran : frekuensi pengeluaran pelanggan
- frekuensi_pemasukan : frekuensi pemasukan pelanggan

EDA

- Banyak pelanggan yang tidak memiliki hutang dan tidak melakukan transaksi.
- Mayoritas dari pelanggan adalah pengusaha dan sudah berkeluarga.
- Mayoritas dari pelanggan memiliki produk 3 dan tidak memiliki produk 5.
- Jumlah pelanggan yang tidak membeli produk 7 jauh lebih banyak dari pelanggan yang membeli produk 7.

Data Preparation

Missing Value Handling & Data Cleaning

Kolom yang memiliki data kosong lebih dari 50% dihapus. Selanjutnya melihat proporsi target, ternyata data yang didapat termasuk imbalance. Dilakukan undersampling untuk menyeimbangkan proporsi target. Data yang telah mengalami undersampling dipisahkan berdasarkan kolomnya, yaitu kolom numerik, kolom kategorik, dan kolom kategorik biner. Data kosong pada kolom numerik dilakukan imputasi dengan nilai mediannya. Data kosong pada kolom kategorik dan kolom kategorik biner dilakukan imputasi dengan nilai yang paling banyak muncul.

Feature Engineering

Feature Transformation

Transformasi yang dilakukan pada kolom numerik berupa scaling menggunakan robust scaler. Pada kolom kategorik, dilakukan encoding dengan one hot encoding. Kolom kategorik biner hanya mengubah nilai pada kolom ‘gender’, nilai ‘F’ menjadi ‘1’ dan nilai ‘M’ menjadi ‘0’. Selanjutnya data disatukan kembali menjadi sebuah dataframe.

Feature Selection

Seleksi fitur dilakukan menggunakan selectkbest. Diambil 50 fitur pertama dengan asumsi 50 fitur ini tetap dapat mewakili informasi dari data secara keseluruhan namun dengan jumlah yang lebih sedikit dari fitur aslinya.

Modeling

Modeling

Model yang dilipilih untuk data ini adalah random forest dan support vector machine.

Hyperparameter Optimization & Cross Validation

Random Forest

Pada model random forest, parameter yang dituning antara lain: max_depth, max_features, max_samples, min_samples_leaf, min_samples_split, dan n_estimators.

```
In [37]: gscv.best_params_

Out[37]: {'max_depth': 4,
          'max_features': 35,
          'max_samples': 0.7,
          'min_samples_leaf': 20,
          'min_samples_split': 50,
          'n_estimators': 45}
```

Hyperparameter Optimization & Cross Validation

Support Vector Machine

Pada model svm, parameter yang dituning antara lain: kernel, gamma (kernel rbf), degree (kernel poly), dan penalty.

In [46]:
```
# membuat model svm yang telah dioptimasi
svc_optimized = SVC(kernel='rbf', gamma=0.1, C=0.1)
svc_optimized.fit(X_train_selected, y_train_rus)
```

Out[46]:
```
SVC(C=0.1, gamma=0.1)
```

Perhitungan Performance

Random Forest

random forest default

	precision	recall	f1-score	support
1	0.94	0.95	0.95	1195
0	0.95	0.94	0.95	1195
accuracy			0.95	2390
macro avg	0.95	0.95	0.95	2390
weighted avg	0.95	0.95	0.95	2390

	precision	recall	f1-score	support
1	0.14	0.58	0.22	299
0	0.92	0.56	0.69	2466
accuracy			0.56	2765
macro avg	0.53	0.57	0.46	2765
weighted avg	0.83	0.56	0.64	2765

random forest optmized

	precision	recall	f1-score	support
1	0.57	0.90	0.70	1195
0	0.76	0.31	0.44	1195
accuracy			0.61	2390
macro avg	0.66	0.61	0.57	2390
weighted avg	0.66	0.61	0.57	2390

	precision	recall	f1-score	support
1	0.13	0.86	0.22	299
0	0.95	0.29	0.45	2466
accuracy			0.36	2765
macro avg	0.54	0.58	0.34	2765
weighted avg	0.86	0.36	0.42	2765

Perhitungan Performance

Support Vector Machine

svm default

	precision	recall	f1-score	support
1	0.59	0.06	0.11	1195
0	0.50	0.96	0.66	1195
accuracy			0.51	2390
macro avg	0.55	0.51	0.38	2390
weighted avg	0.55	0.51	0.38	2390

	precision	recall	f1-score	support
1	0.07	0.03	0.05	299
0	0.89	0.95	0.92	2466
accuracy			0.85	2765
macro avg	0.48	0.49	0.48	2765
weighted avg	0.80	0.85	0.82	2765

svm optmized

	precision	recall	f1-score	support
1	0.55	0.51	0.53	1195
0	0.54	0.59	0.56	1195
accuracy			0.55	2390
macro avg	0.55	0.55	0.55	2390
weighted avg	0.55	0.55	0.55	2390

	precision	recall	f1-score	support
1	0.13	0.51	0.20	299
0	0.91	0.57	0.70	2466
accuracy			0.56	2765
macro avg	0.52	0.54	0.45	2765
weighted avg	0.82	0.56	0.65	2765

Evaluasi

Backtesting

Berdasarkan perhitungan performansi, untuk backtesting digunakan model random forest dufault.