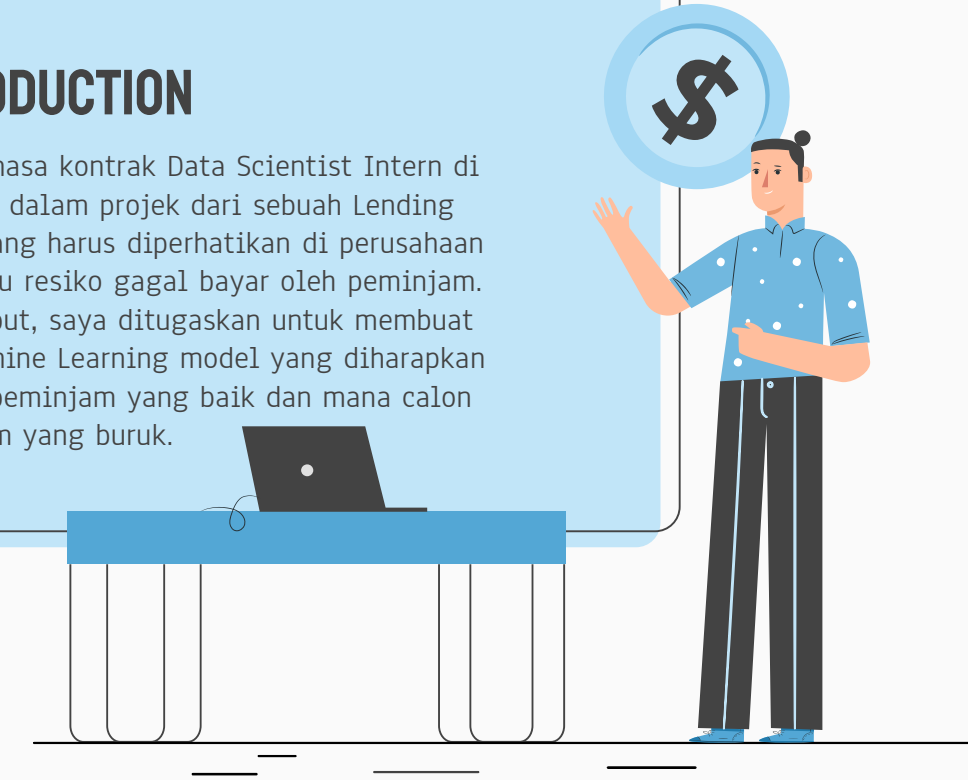
CREDIT RISK MODELLING

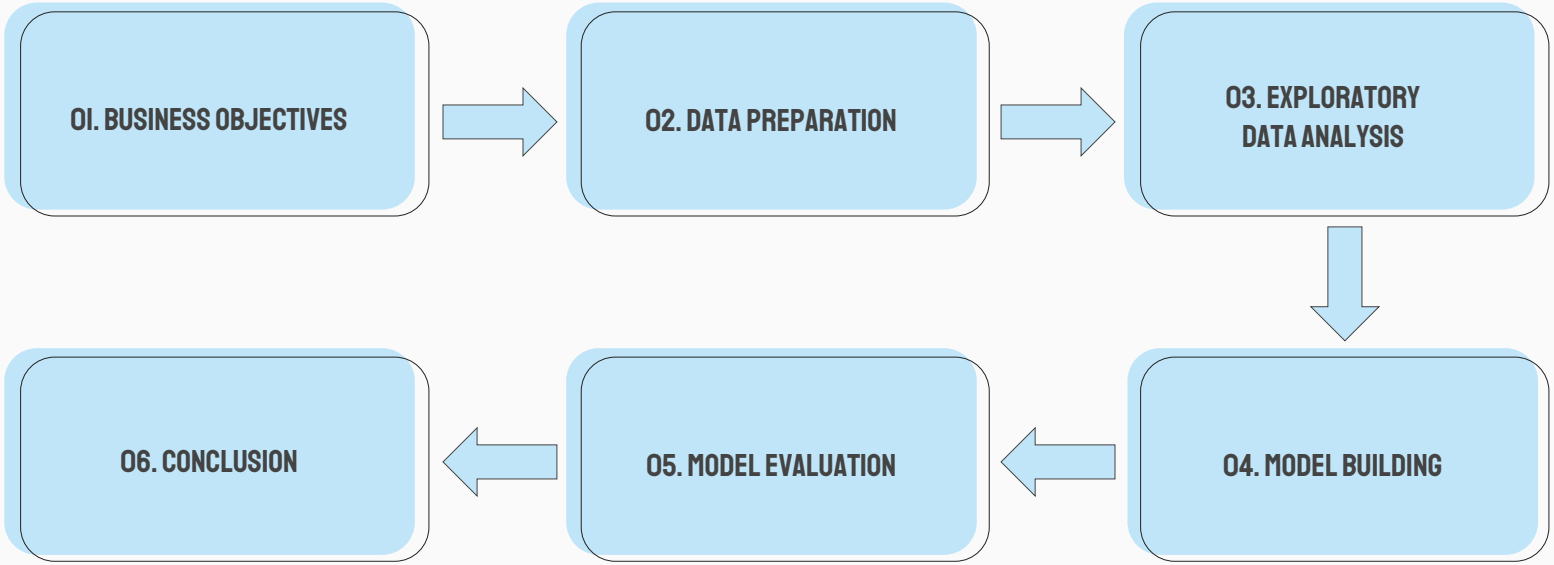
Loan Approval Classification

INTRODUCTION

Sebagai tugas akhir saya dari masa kontrak Data Scientist Intern di ID/X Partners, saya dilibatkan dalam projek dari sebuah Lending Company. Satu hal terpenting yang harus diperhatikan di perusahaan lending adalah Credit Risk, yaitu resiko gagal bayar oleh peminjam. Untuk meminimalisir hal tersebut, saya ditugaskan untuk membuat end-to-end solution berupa Machine Learning model yang diharapkan bisa memprediksi mana calon peminjam yang baik dan mana calon peminjam yang buruk.



STEP-BY-STEP


01. BUSINESS OBJECTIVES
02. DATA PREPARATION
03. EXPLORATORY DATA ANALYSIS
04. MODEL BUILDING
05. MODEL EVALUATION
06. CONCLUSION

01. BUSINESS OBJECTIVES

BUSINESS OBJECTIVES & METRICS

MINIMIZE RISKS

Meminimalisir resiko kerugian perusahaan dengan mengklasifikasikan calon debitur yang memiliki kemungkinan gagal membayar hutang (Recall).

MAXIMIZE RETURNS

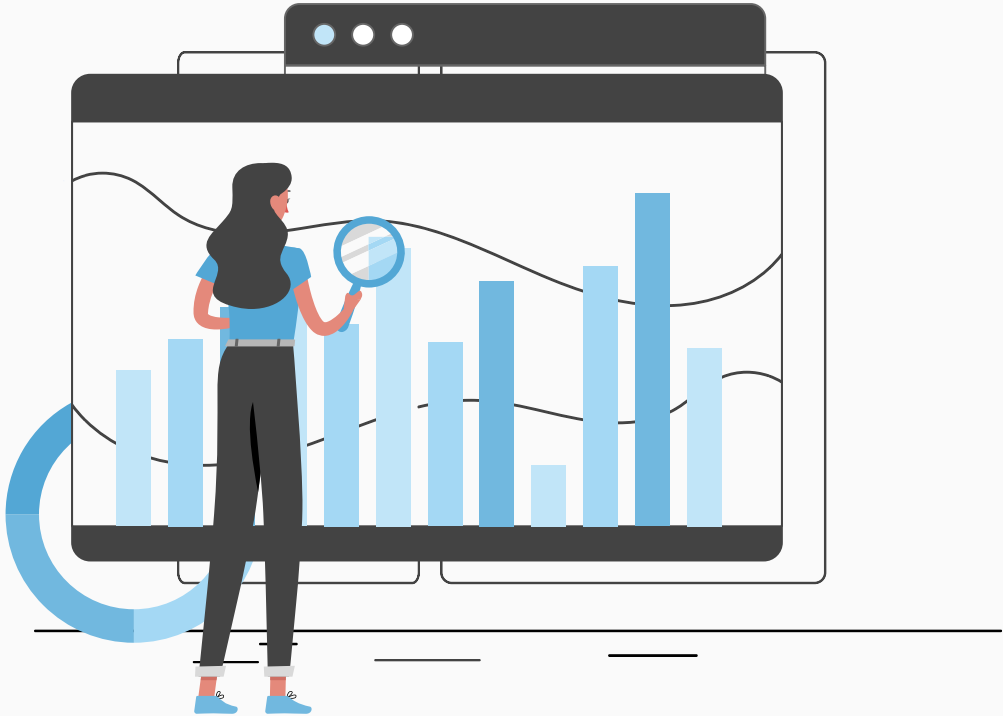
Memaksimalkan keuntungan perusahaan dengan mengklasifikasikan calon debitur yang memiliki kemungkinan sukses membayar hutang (Precision).

02. DATA PREPARATION

DATA PREPARATION

1. Data memiliki 466285 rows, dan 75 columns.
2. Tidak terdapat observasi duplikat pada data.
3. Membuang kolom - kolom yang merupakan identifier.
4. Membuang kolom - kolom yang memiliki missing values 100%.
5. Mentransform kolom “loan_status” yang merupakan variabel target menjadi 1 untuk “pinjaman buruk” dan 0 untuk “pinjaman baik”.

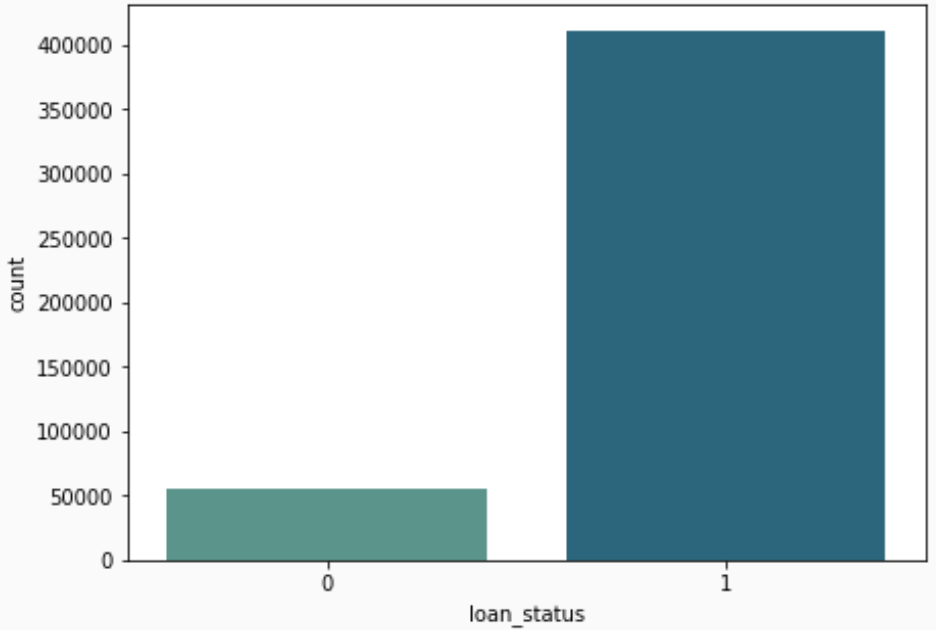
03.
EXPLORATORY DATA ANALYSIS



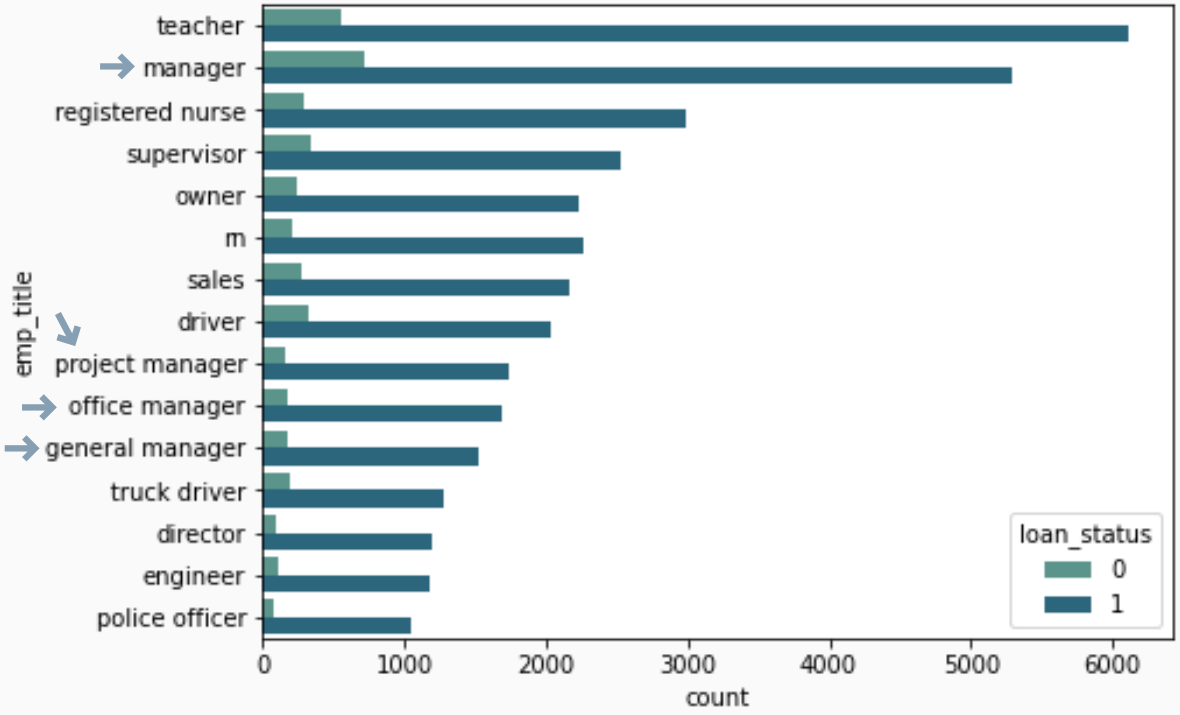
TARGET VARIABLE

```
[ ]  df['loan_status'].value_counts(normalize=True)

     1    0.881334
     0    0.118666
     Name: loan_status, dtype: float64
```

Saat ini perusahaan masih gagal mengklasifikasikan calon peminjam buruk sebesar 11.8% dari total debitur.



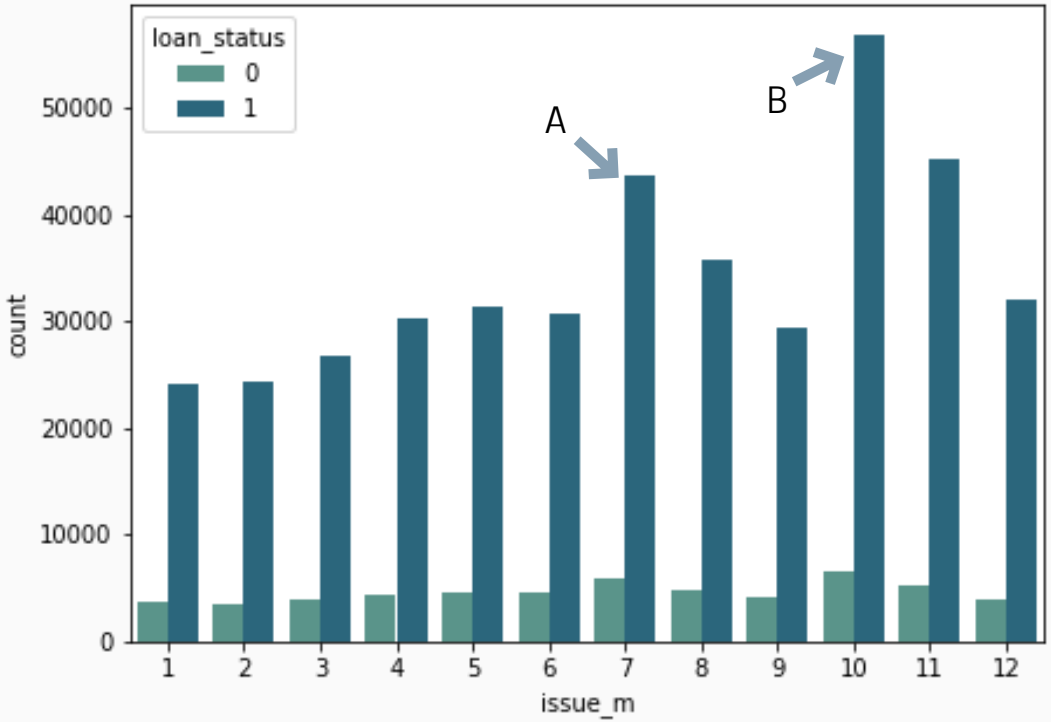
JOB TITLE VS LOAN STATUS



1. Guru merupakan profesi yang paling banyak menjadi debitur.
2. Namun jika kita lihat lebih dekat, managerlah yang paling banyak menjadi debitur.

3. Hal yang menarik adalah presentase seorang manager menjadi peminjam yang buruk lebih besar daripada guru.

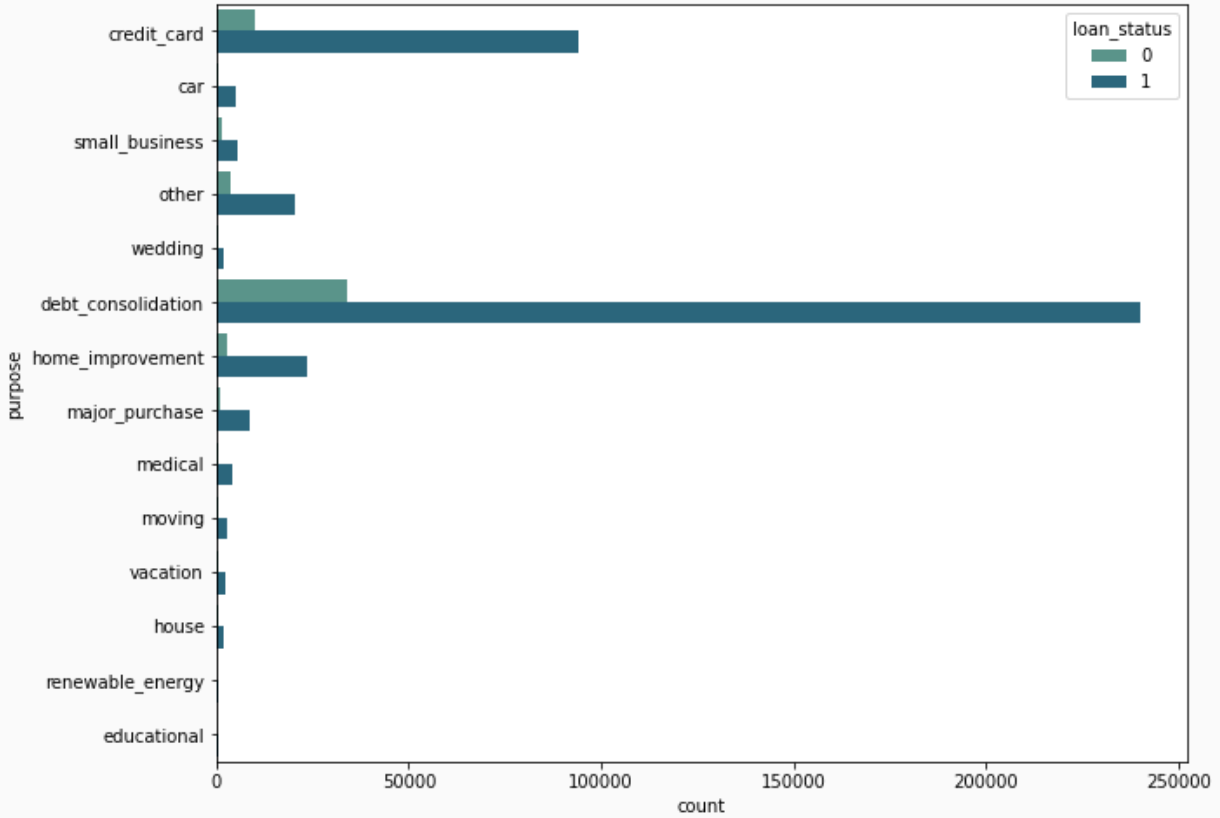
SEASONALITY TREND



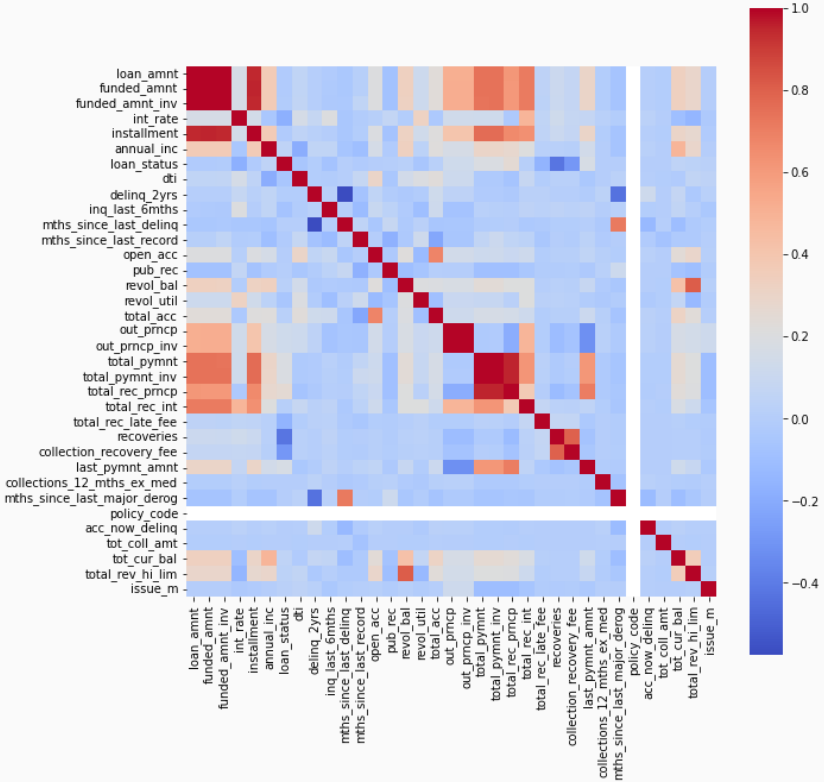
1. Terdapat 2 trend seasonality bulan peminjaman.
2. Trend pertama yaitu A merupakan bulan Juli, hal ini bertepatan dengan hari libur kemerdekaan Amerika Serikat.
3. Trend kedua yaitu bulan Oktober-November-Desember, hal ini bertepatan dengan Haloween pada bulan Oktober, Natal dan Tahun baru pada bulan Desember.
4. Tidak terdapat trend di bulan tertentu terhadap peminjam yang buruk. Rata-rata credit risk tiap bulan berada di angka -+ 10%.

LOAN PURPOSE

1. Tujuan seseorang meminjam uang yang paling banyak adalah melunasi hutang lain dan kartu kredit.
2. Ironisnya, kedua hal tersebut merupakan penyumbang keuntungan & kerugian terbesar untuk perusahaan.
3. Bisa dibilingan hal tersebut merupakan “High risk, high reward” untuk perusahaan.

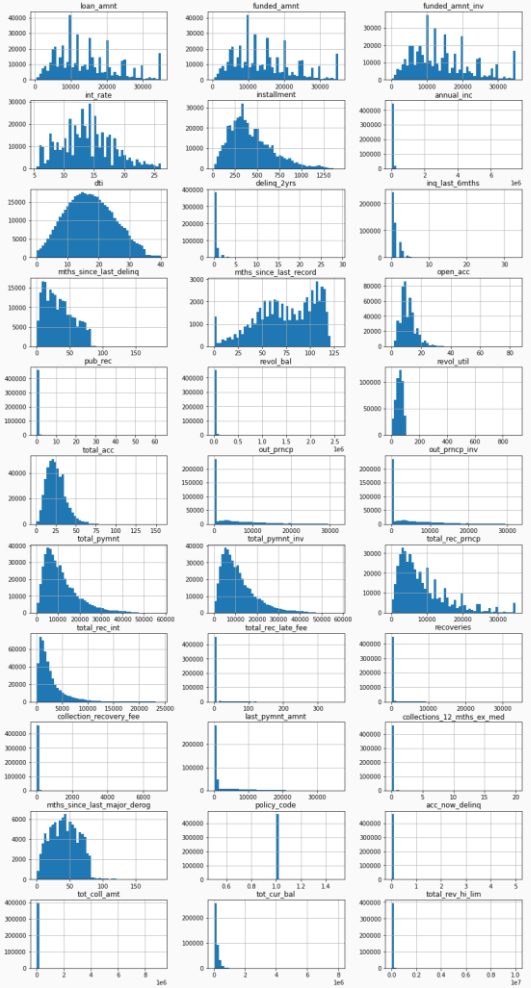


MULTICOLLINEARITY



1. Terdapat beberapa kolom yang memiliki korelasi antar variabel bernilai 1. Yang artinya kedua kolom tersebut memiliki informasi yang sama/identik. Kolom tersebut perlu kita buang agar model yang akan kita buat bisa bekerja dengan maksimal.

2. Bisa kita lihat juga kolom “policy_code” tidak memiliki nilai apapun kita juga akan membuang kolom tersebut.



NUMERICAL FEATURES DISTRIBUTION

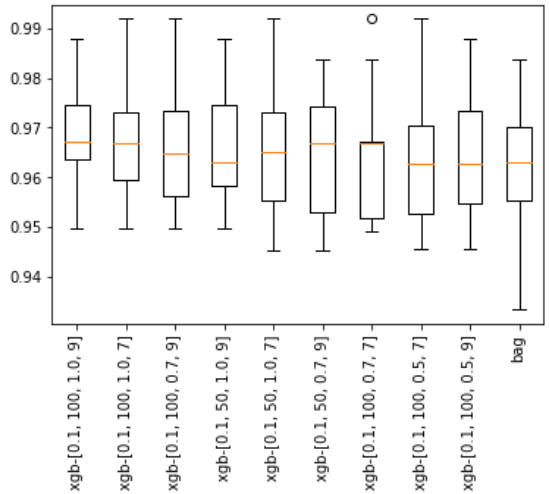
1. Melihat distribusi fitur numerical sekilas, kita bisa melihat banyak terdapat fitur yang mempunyai ekor di kanan yang artinya right-skewed.
2. Dengan informasi tersebut kita bisa menyimpulkan untuk mengunakan median sebagai angka untuk mengimpute missing values.

3. Terdapat beberapa variabel yang memiliki nilai ekstrim/outliers, untuk saat ini kita akan mengabaikan hal itu dan melihat performa model.

04. MODEL BUILDING

SPOT-CHECK ALGORITHM WITH SAMPLE DATA

```
Rank=1, Name=xgb-[0.1, 100, 1.0, 9], Score=0.969 (+/- 0.011)
Rank=2, Name=xgb-[0.1, 100, 1.0, 7], Score=0.968 (+/- 0.012)
Rank=3, Name=xgb-[0.1, 100, 0.7, 9], Score=0.966 (+/- 0.013)
Rank=4, Name=xgb-[0.1, 50, 1.0, 9], Score=0.966 (+/- 0.013)
Rank=5, Name=xgb-[0.1, 50, 1.0, 7], Score=0.966 (+/- 0.013)
Rank=6, Name=xgb-[0.1, 50, 0.7, 9], Score=0.965 (+/- 0.014)
Rank=7, Name=xgb-[0.1, 100, 0.7, 7], Score=0.965 (+/- 0.014)
Rank=8, Name=xgb-[0.1, 100, 0.5, 7], Score=0.964 (+/- 0.013)
Rank=9, Name=xgb-[0.1, 100, 0.5, 9], Score=0.964 (+/- 0.012)
Rank=10, Name=bag, Score=0.962 (+/- 0.014)
```



1. Pertama, kita hanya menggunakan 10.000 data pertama untuk mencoba beberapa algoritma menggunakan 3 CV-Validation.

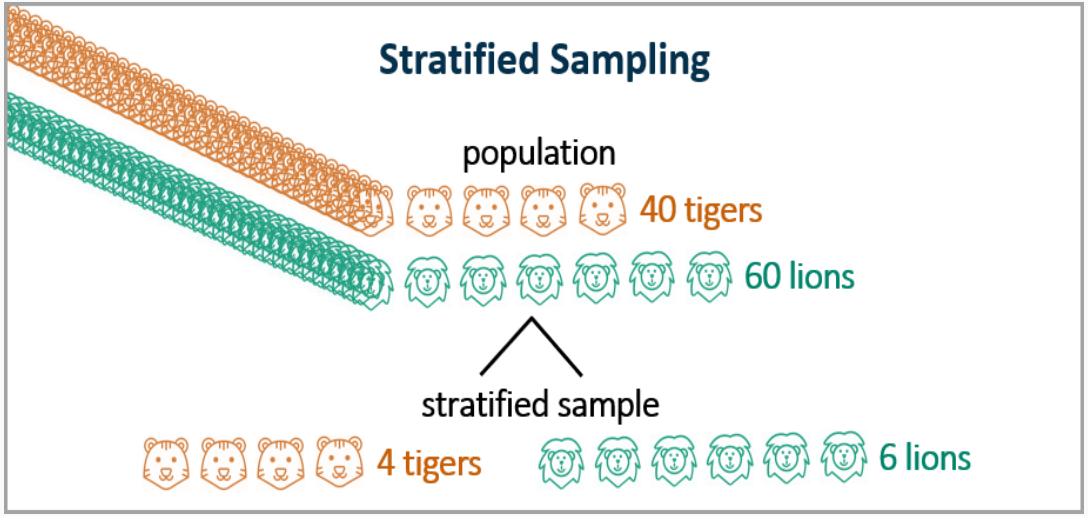
2. Terlihat dari plot di samping algoritma XGBoost dan BaggingClassifier memiliki F-1 Score tertinggi dibandigkan algoritma-algoritma yang lain.

SPLIT DATASET INTO TRAIN & TEST SET

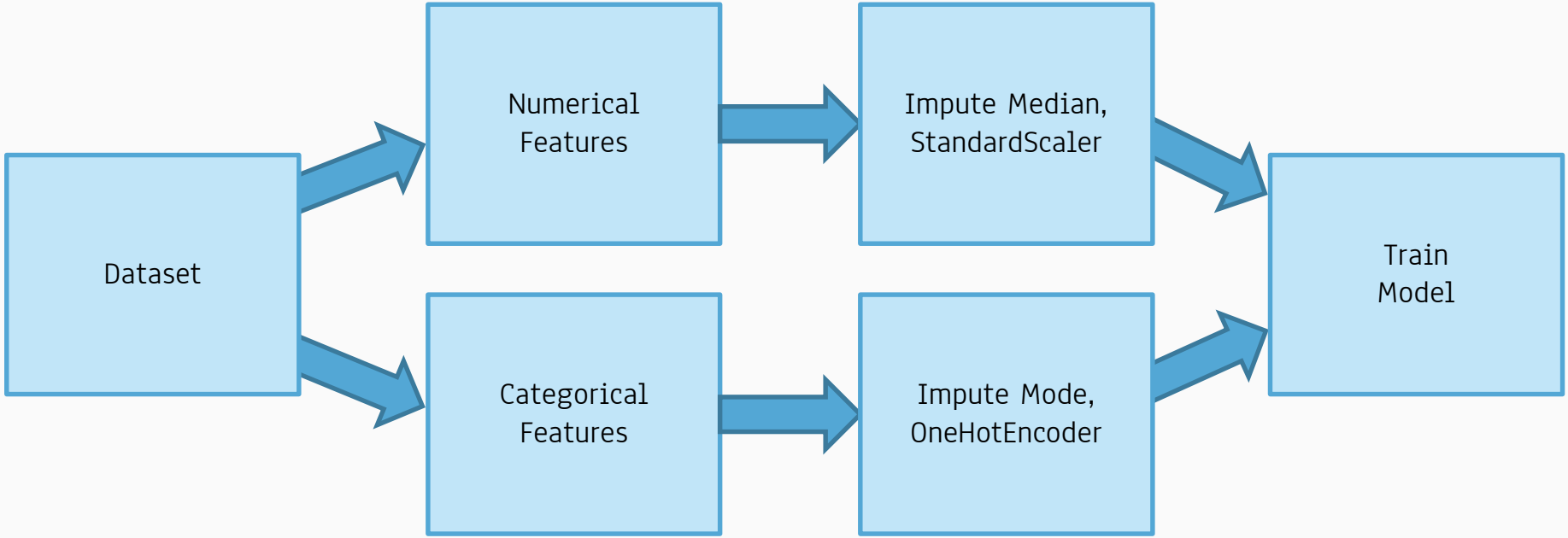
```
[83] from sklearn.model_selection import train_test_split

     X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, stratify=y, random_state=42)
```

1. Setelah melakukan spot-check kita akan melakukan training pada full dataset.

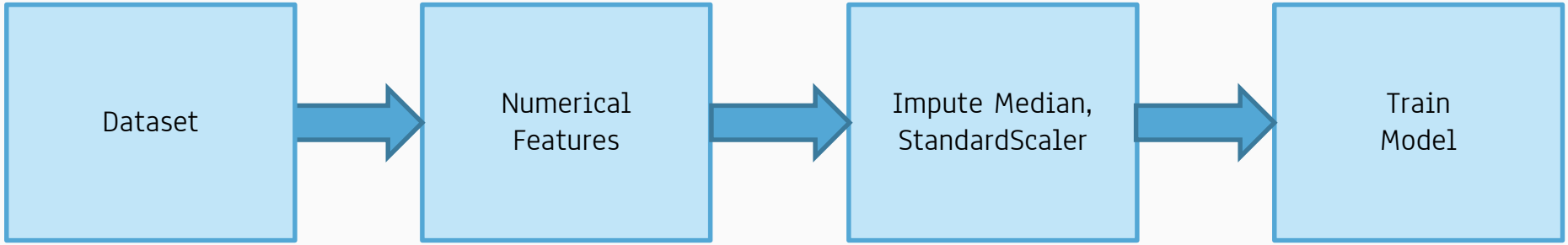
2. Hal yang pertama kita lakukan adalah memecah dataset menjadi train set dan test set, agar tidak terjadi target leaking.



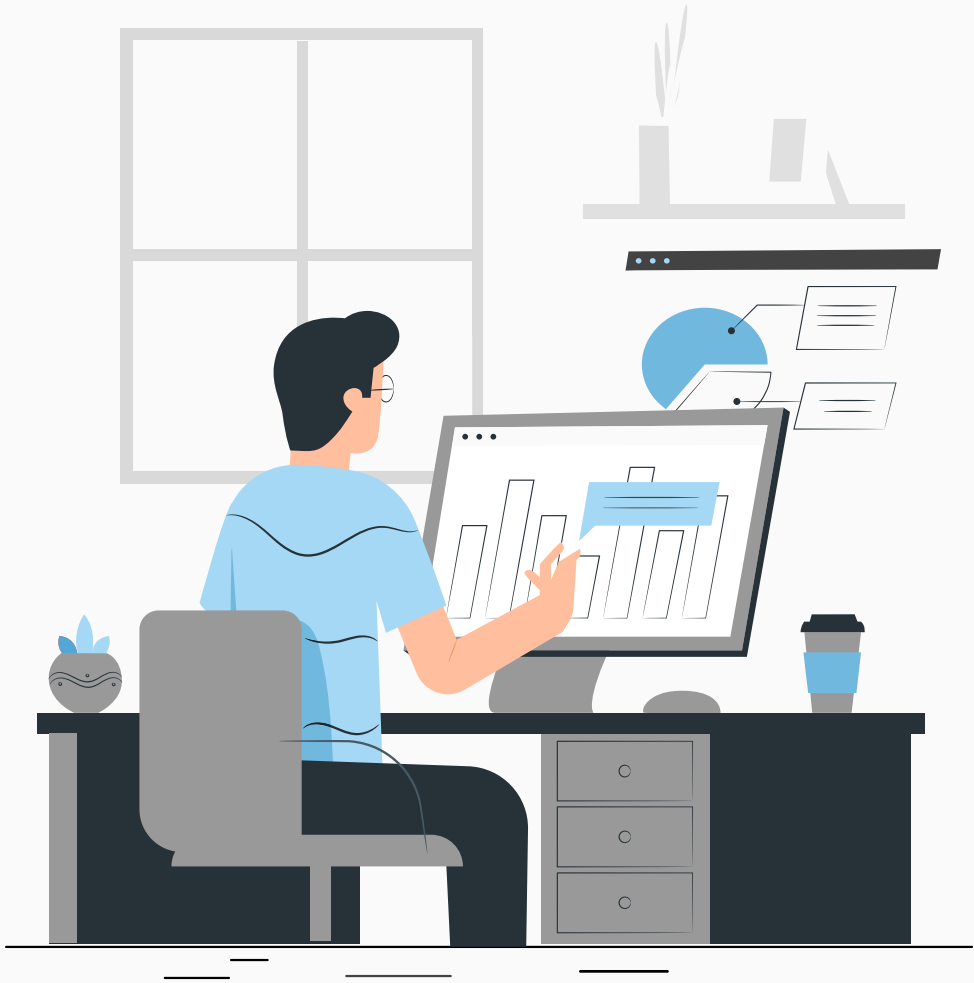
PIPELINE ALL FEATURES



PIPELINE NUMERICAL FEATURES ONLY



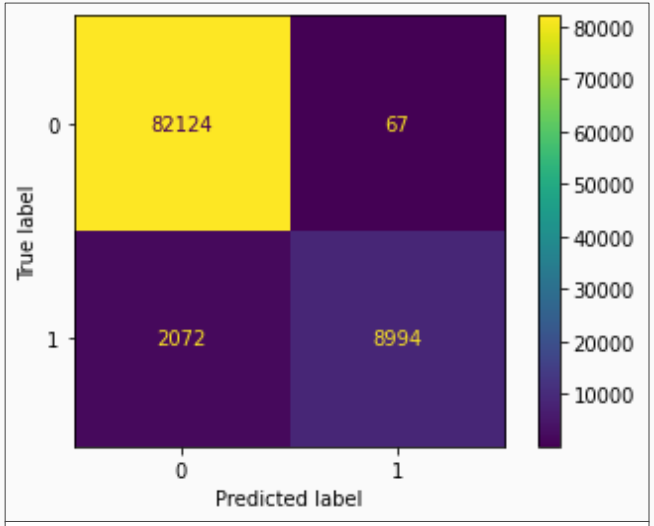
05. MODEL EVALUATION



ALL FEATURES VS NUMERICAL FEATURES ONLY

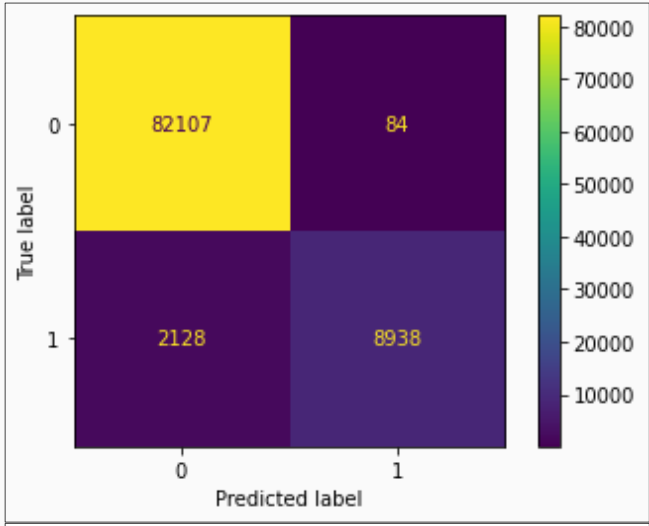
1. Tidak terdapat penurunan score yang signifikan setelah membuang fitur kategorikal yang berjumlah -+ 70.
2. Hal ini menguntungkan karena mengurangi kompleksitas model dan membuat model lebih generalisir.

All Features (Numerical + 70 Categorical Features)



	precision	recall	f1-score	support
0	0.98	1.00	0.99	82191
1	0.99	0.81	0.89	11066
accuracy			0.98	93257
macro avg	0.98	0.91	0.94	93257
weighted avg	0.98	0.98	0.98	93257

Numerical Features Only

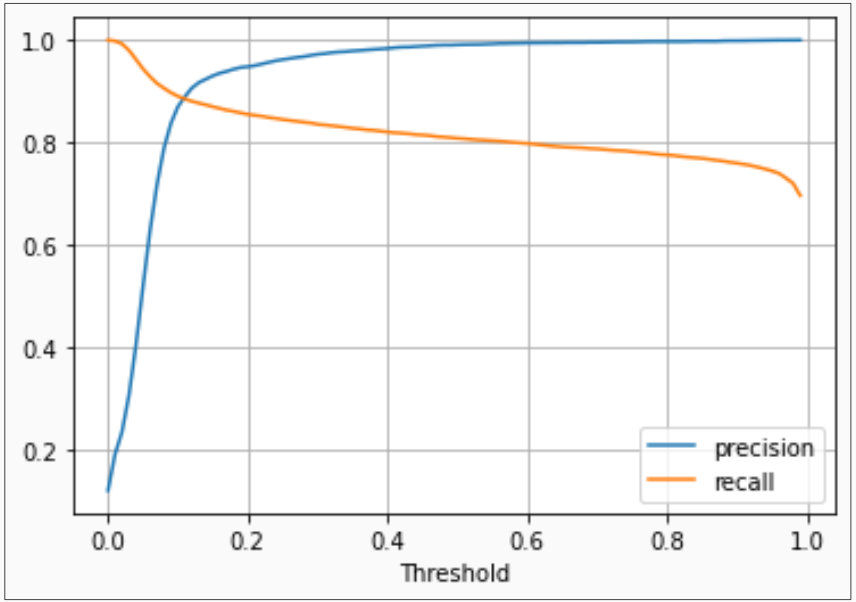


	precision	recall	f1-score	support
0	0.97	1.00	0.99	82191
1	0.99	0.81	0.89	11066
accuracy			0.98	93257
macro avg	0.98	0.90	0.94	93257
weighted avg	0.98	0.98	0.98	93257

PRECISION RECALL TRADE-OFF

Konteks

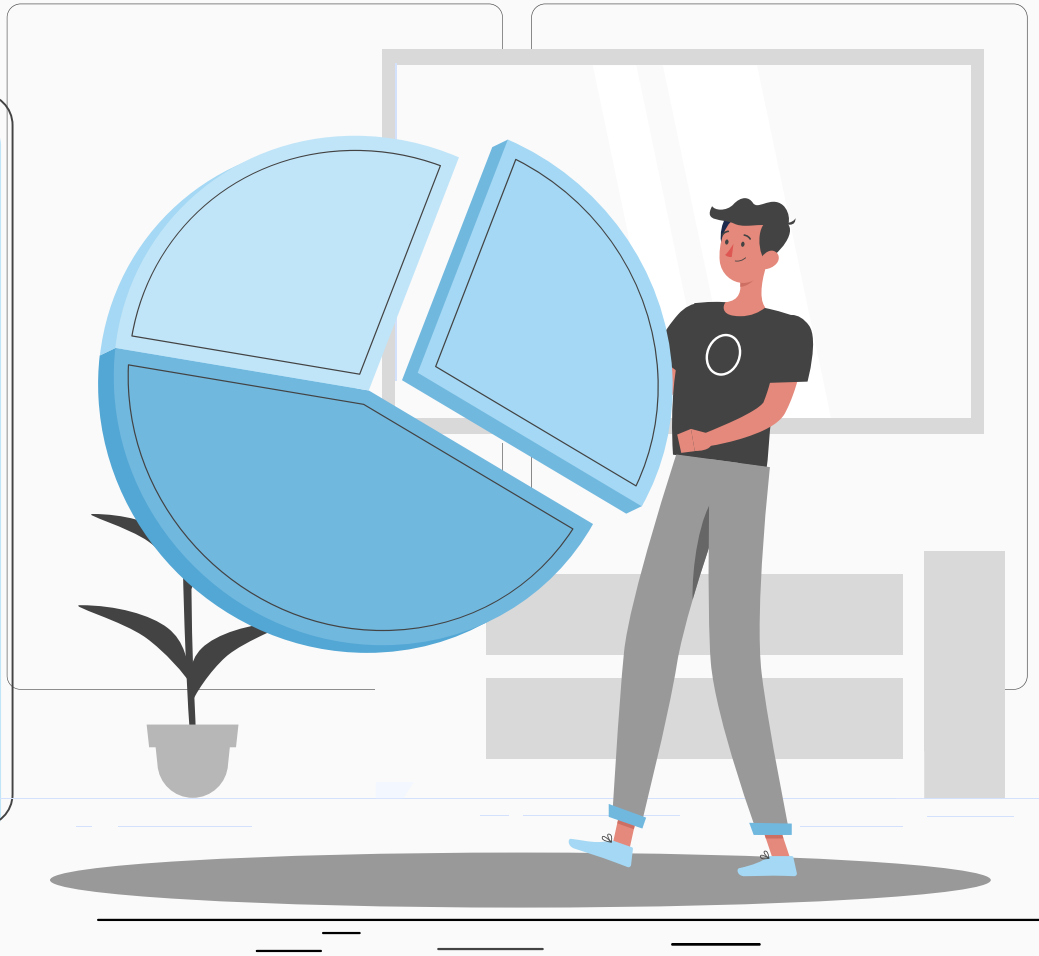
Saat ini recall perusahaan berada di angka 88.1%. Diharapkan model yang kita buat melebihi angka terserbut dalam memprediksi calon pemijam buruk.



Dari tabel recall vs precision tersebut, angka yang paling masuk akal adalah diantara 0.08 - 0.10. Karena precision masih diatas 75% dan recall diatas 88.1%.

	precision	recall	threshold
0	0.118661	1.000000	0.00
1	0.189062	0.998102	0.01
2	0.233053	0.992590	0.02
3	0.305130	0.979939	0.03
4	0.403151	0.962046	0.04
5	0.517055	0.943792	0.05
6	0.625822	0.928610	0.06
7	0.716773	0.915236	0.07
8	0.787926	0.905838	0.08
9	0.836931	0.896982	0.09
10	0.869657	0.889933	0.10

06. CONCLUSION



CONCLUSION

1. Model yang kita buat berhasil memenuhi standar perusahaan dengan recall diatas 88.1% untuk memprediksi calon peminjam buruk (XGBoost = learning_rate=0.1, n_estimators=100, subsample=1.0, max_depth=9)
2. Tidak ada data pembanding untuk menghitung presentase calon peminjam baik yang ditolak.
3. Fitur categorical hanya berkontribusi menaikkan F1-score sebesar 0.4% meskipun terdapat -+ 70 fitur baru setelah One Hot Encoding.
4. Fitur numerical yang kita gunakan dalam membuat model menggunakan median untuk mengimpute missing value dikarenakan terdapat banyak outliers pada data.
5. Score model mungkin masih bisa meningkat dengan mengolah data outliers (membuang outliers, mengganti outliers dengan angka maksimum setelah ouliers dibuang, dll), mentransform distribusi fitur, dll.
6. Feature importance & selection bisa dilakukan dengan harapan model lebih generalisir dan tidak overfit.
7. Dikarenakan kita melakukan spot-check algorithm pada data sample (dikarenakan time and resource constraints) mungkin saja model yang kita buat masih belum maksimal, langkah yang bisa dilakukan adalah spot-check dan hyperparameter tuning pada full dataset.

THANKS

Does anyone have any questions?

raynaldydk10@gmail.com
+62877-1434-2574