Prediksi Target Pelanggan J Taxi

Tim ElsiX
Atmavidya Virananda
M. Sammy Ivan K
J. J. Billie C.

Content

1. Business Understanding
2. Exploratory Data Analysis
3. Preprocessing
4. Feature Engineering
5. Modelling
6. Evaluation

1.

Business Understanding

Peningkatan Layanan J Taxi

- Sebuah perusahaan taksi seperti J Taxi haruslah mengedepankan *customer satisfaction* sebagai satu tonggak utamanya dalam *value delivery*. Dengan banyaknya data yang terdapat di era sekarang ini, serta fakta bahwa J Taxi telah beroperasi selama 1 tahun, analitika data dan pendekatan *data science* dapat menjadi *approach* yang baik.

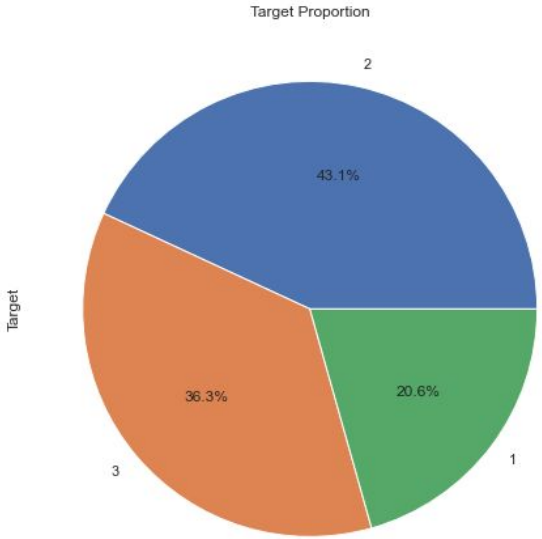
- Dengan memanfaatkan data historis *trip* yang memiliki sedikit informasi terkait pelanggan yang ada di *trip* tersebut, pengelompokkan terhadap pelanggan dapat memberikan arahan bagi J Taxi untuk dapat:
  - Menargetkan layanannya terhadap tendensi
  - Meningkatkan *customer satisfaction* dengan layanan yang tepat
  - Membuat efisiensi dan efektivitas layanan menjadi lebih tinggi dalam segi biaya dan *value delivery*

2.

Exploratory Data Analysis

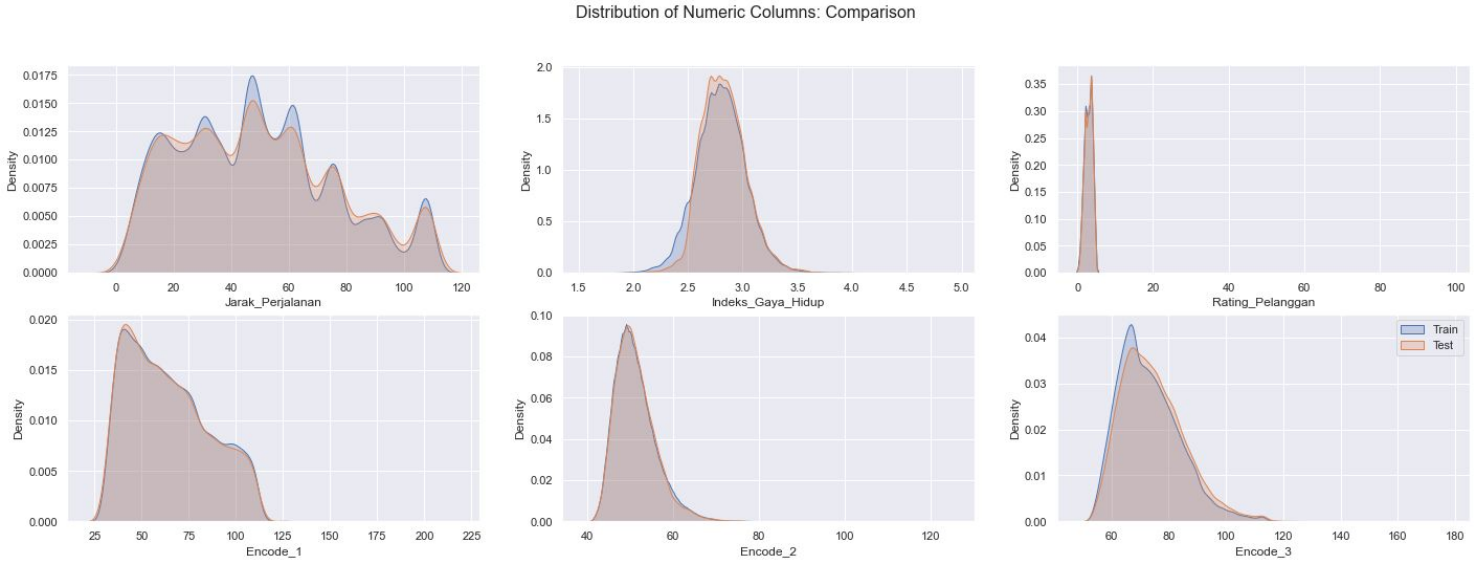
EDA: Proporsi *Target Variable*

- Proporsi *target* menjadi penting untuk diketahui karena dataset yang ***imbalanced*** akan mengakibatkan distorsi prediksi kepada model nantinya.
- Ditemukan bahwa proporsi dari ketiga jenis *target* adalah sebagaimana terlihat di *pie chart* di samping,
  - 1 - 20.6%
  - 2 - 43.1%
  - 3 - 36.3%
- Ditetapkan bahwa *dataset* ini **tidak *imbalanced***.



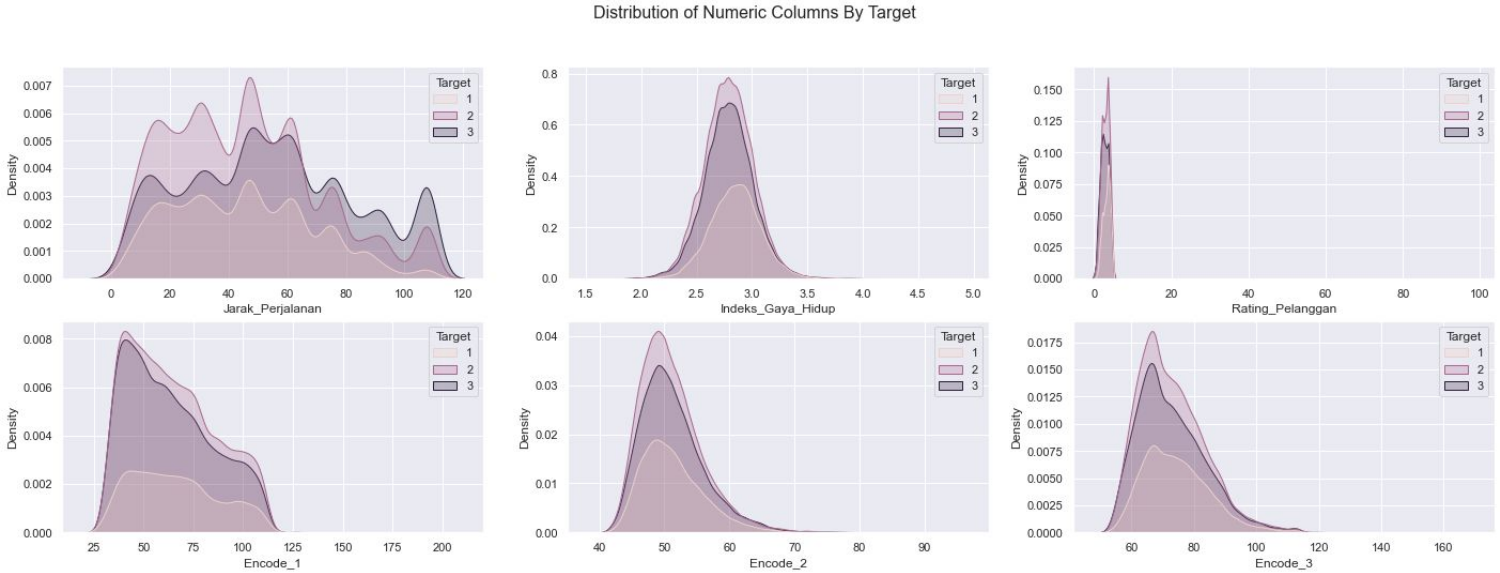
EDA: Analisis Distribusi Fitur Antara *Train* & *Test*

- Dengan menggunakan ***kdeplot***, ditemukan bahwa distribusi dari kolom-kolom numerik di bawah serupa untuk *train* dan *test*.
- Ditemukan indikasi *outlier* pada *dataset*, namun belum terlihat pada *train* atau *test*. Penelitian lebih lanjut menemukan bahwa *dataset train* yang memiliki *outlier*.



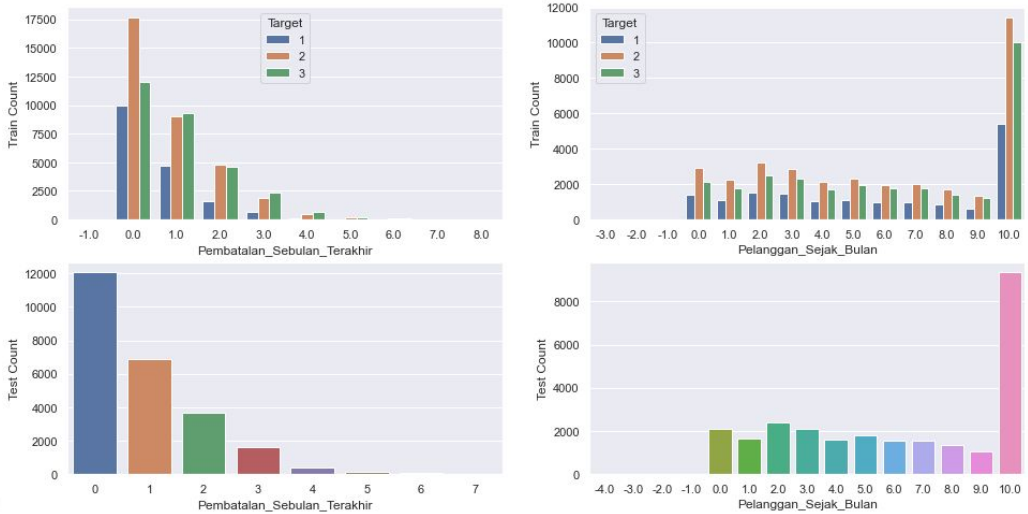
EDA: Analisis Distribusi Fitur Antara *Train* & *Test*

- Dengan menggunakan ***kdeplot***, ditemukan bahwa distribusi dari kolom-kolom numerik di bawah berdasarkan *target* memiliki distribusi yang cukup serupa untuk tiap *target*.
- Pelanggan *target* "1" memiliki kecenderungan lebih rendah untuk bepergian jarak di atas 100 dibandingkan *target* "2" dan "3"



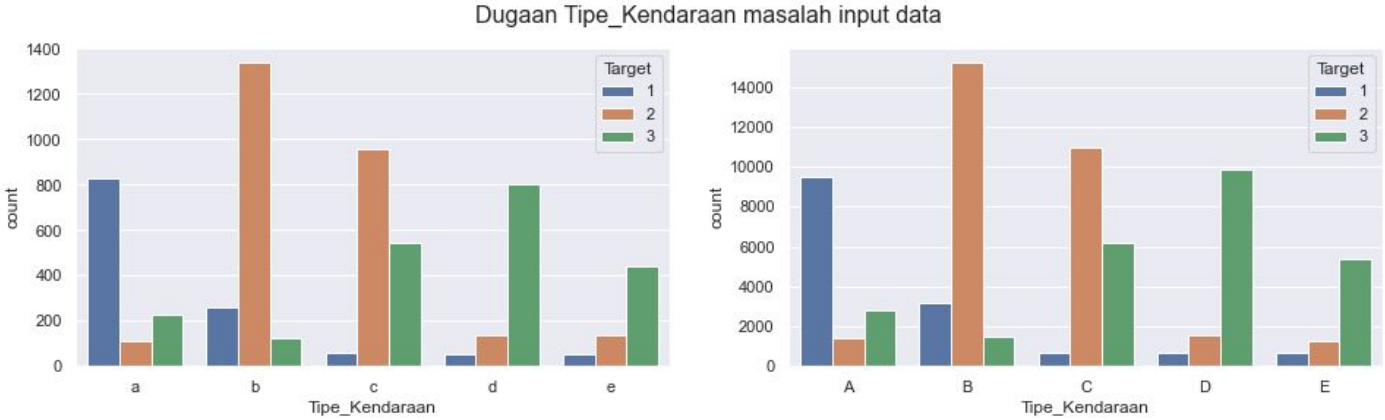
EDA: Analisis Distribusi Fitur Antara *Train* & *Test*

- Dua fitur dari fitur-fitur numerik yang ada menunjukkan nilai-nilai *integer* yang **tidak sepenuhnya kontinyu**. Oleh karena itu, digunakan ***countplot*** untuk memvisualisasikan distribusinya.
- Ditemukan bahwa terdapat nilai **"Pembatalan_Sebulan_Terakhir"** dan **"Pelanggan_Sejak_Bulan"** yang tidak masuk akal. Selain hal tersebut, distribusi kedua fitur serupa.
- Ditemukan bahwa proporsi target sesuai dengan proporsi target keseluruhan.
- Kebanyakan pelanggan sudah menggunakan jasa J Taxi dari 10 bulan yang lalu.

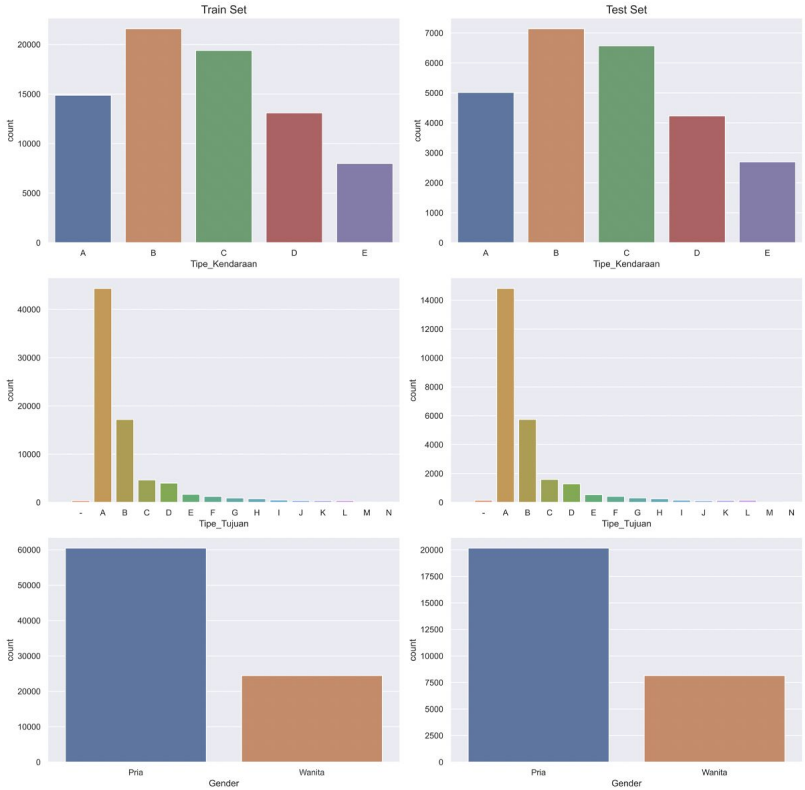


EDA: Analisis Distribusi Fitur Antara *Train* & *Test*

- Sebelum melanjutkan ke distribusi kategorikal, terdapat apriori bahwa jenis **“Tipe_Kendaraan” memiliki inkonsistensi,** dimana tipe dengan huruf latin dituliskan secara kapital dan huruf kecil. Dalam hal ini, dicurigai bahwa ini adalah masalah *human error*.
- Maka dibandingkan antara **distribusi huruf kapital dan huruf kecil**. Ditemukan bahwa **distribusi serupa** antara keduanya.
- Pelanggan *target* tipe 1, 2, dan 3 memiliki kecenderungan memakai tipe kendaraan yang berbeda-beda



EDA: Analisis Distribusi Fitur Antara *Train* & *Test*

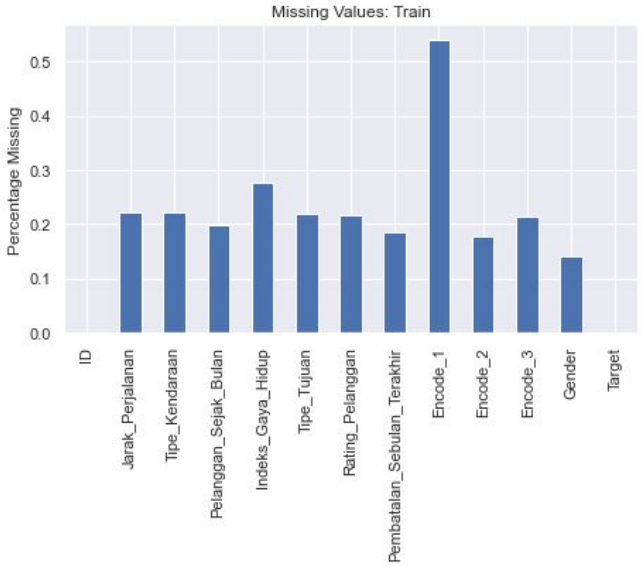


- Distribusi dari 3 fitur kategorikal ditinjau antara *train* dan *test*. Ditemukan bahwa **kedua *dataset* memiliki distribusi** yang serupa.

- *Insights* yang diperoleh:
  - Tipe kendaraan yang paling sering digunakan adalah tipe “B”.
  - Tipe tujuan yang paling sering dituju adalah tipe “A”.
  - Terdapat tipe tujuan dengan nilai “-”
  - Kebanyakan dari pelanggan disini adalah Pria.

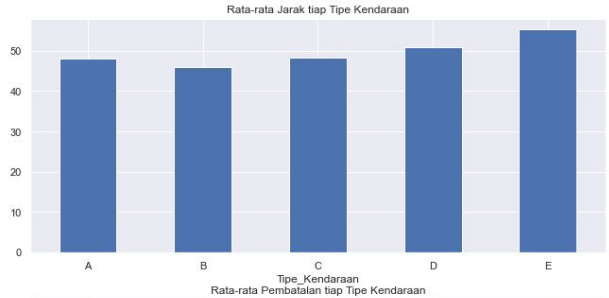
EDA: *Missing Values*

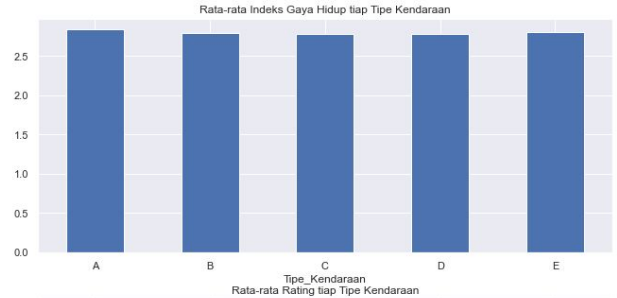
- *Missing values* akan menjadi masalah nantinya apabila tidak diatasi, oleh karena itu dilihat proporsi *missing values* untuk tiap dataset
- Ditemukan bahwa setiap fitur selain “ID” dan *target variable* terdapat *missing values* yang nilainya signifikan
- Khusus untuk fitur **“Encode_1”** terdapat *missing values* yang sangat banyak, yakni **di atas 50%**

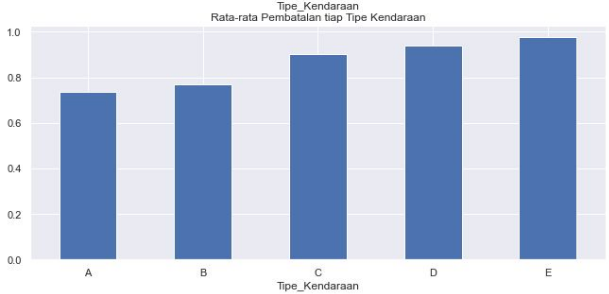


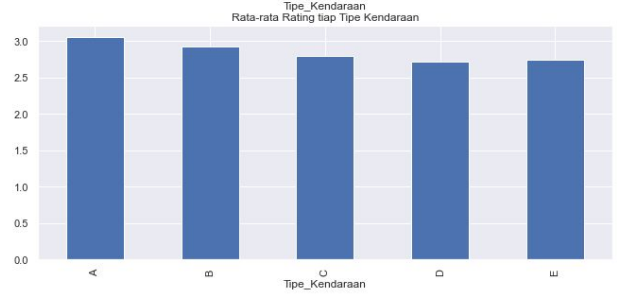
EDA: *Grouped Numerics By* Tipe_Kendaraan

- Karena tipe kendaraan memiliki distribusi *target* yang berbeda-beda, maka akan dilihat fitur lain yang diagregasi menjadi rata-rata terhadap tiap tipe kendaraan.
- Ditemukan bahwa untuk fitur numerik ini tidak ditemukan perbedaan signifikan.



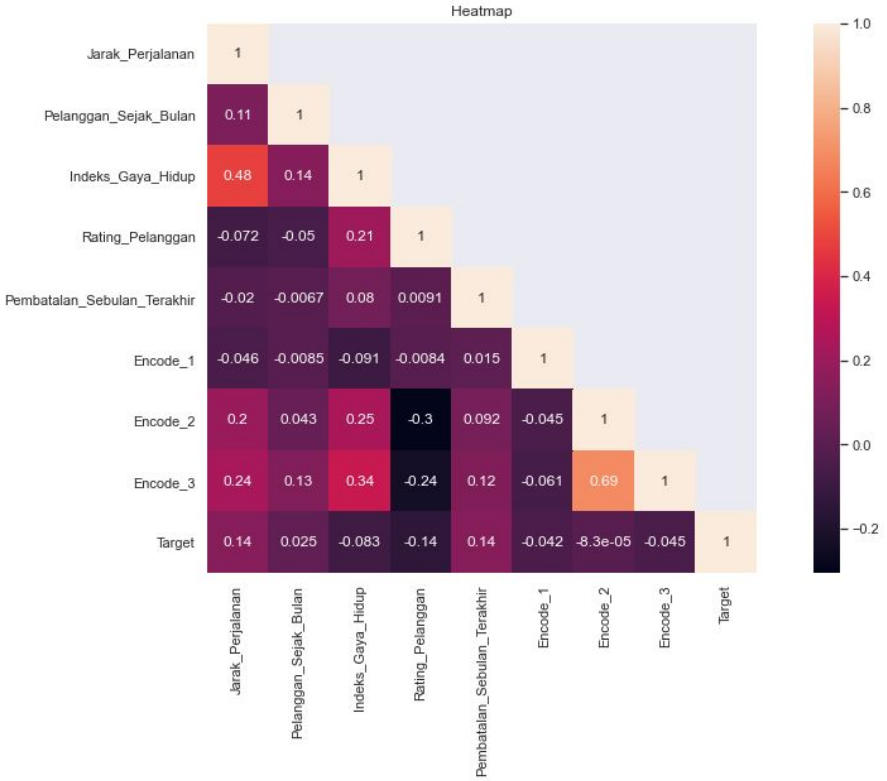






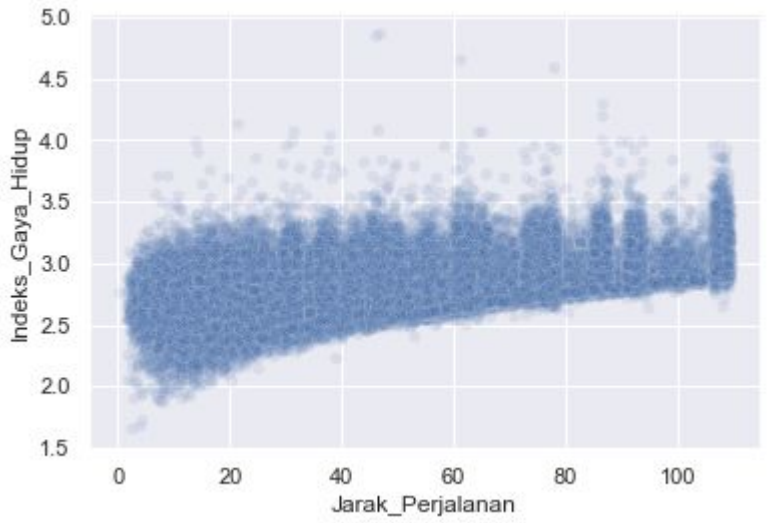
EDA: Korelasi Antar Fitur

- Korelasi antar variabel dapat menjadi apriori utama untuk pengembangan model.
- Ditemukan bahwa hubungan erat ditemukan di antara **Encode_2** dan **Encode_3**
- Terdapat dugaan hubungan antara **Jarak_Perjalanan dengan Indeks_Gaya_Hidup** dengan *pearson correlation coefficient* sebesar 0.48



EDA: Jarak_Perjalanan VS Indeks_Gaya_Hidup

- Berdasarkan korelasi yang ditemukan sebelumnya, digunakan ***scatterplot*** untuk melihat hubungan antara Jarak_Perjalanan dengan Indeks_Gaya_Hidup.
- Dapat dilihat terdapat hubungan yang cukup linear antara kedua fitur sehingga dapat dijadikan acuan untuk pengembangan model.

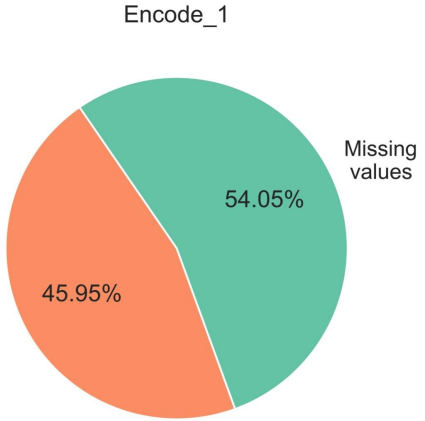


3.
Data Preprocessing

Data Preprocessing: Pembuangan Kolom dan *Outbound Values*

Terdapat dua buah fitur yang dibuang dari dataset, yaitu

- Fitur '**ID**', karena dirasa tidak berpengaruh terhadap variabel target
- Fitur '**Encode_1**', karena memiliki persentase *missing values* sebesar 54%

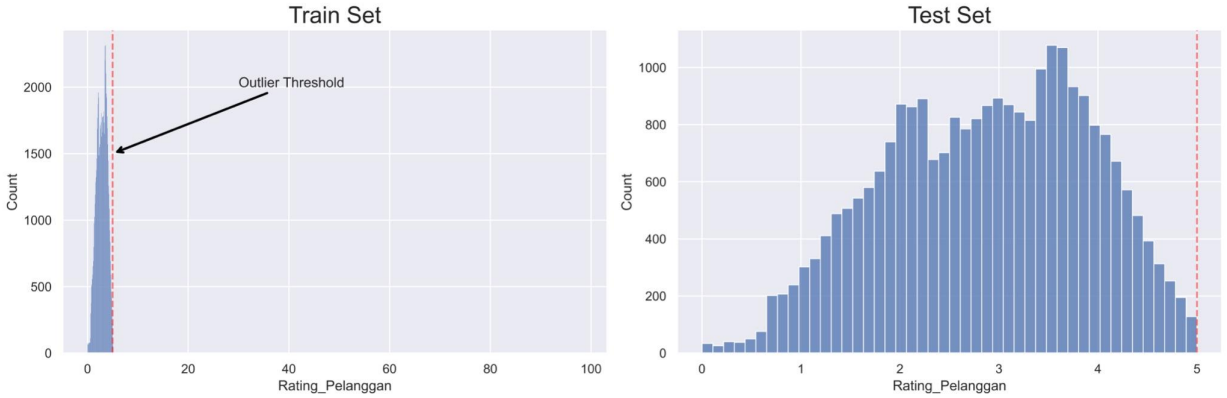


Terdapat baris yang memiliki ***negative values*** pada fitur '**Pelanggan_Sejak_Bulan**' dan '**Pembatalan_Sebulan_Terakhir**'. Diketahui bahwa kedua fitur tersebut tidak dapat bernilai negatif sehingga baris yang bersangkutan dibuang dari dataset.

- **'Pelanggan_Sejak_Bulan'** : 39 baris
- **'Pembatalan_Sebulan_Terakhir':** 2 baris

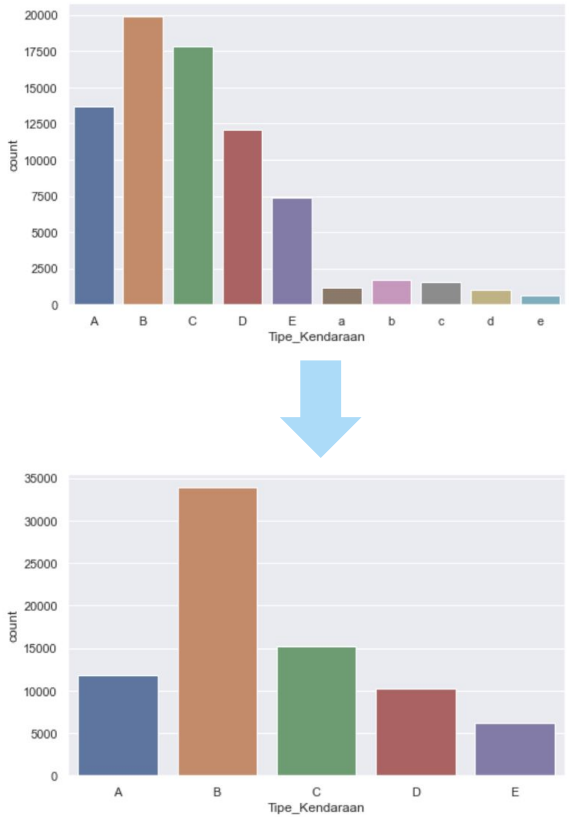
Data Preprocessing: Pembuangan *Outliers*

Pada histogram fitur '**Rating_Pelanggan**' data *train*, dapat terlihat bahwa terdapat *outliers* yang cukup banyak. Nilai pada fitur tersebut memiliki jangkauan nol sampai dengan 98. Sedangkan, jika dilihat pada histogram fitur yang sama di data *test,* jangkauannya hanya dari nol sampai 5. Maka, diputuskan untuk **membuang baris pada data train dengan 'Rating_Pelanggan' di atas 5,** yaitu sebanyak 3 baris.



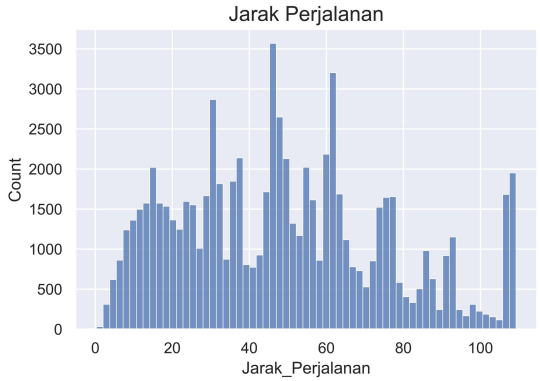
Data Preprocessing: Penggabungan Kategori Tipe_Kendaraan

- Penggabungan kategori **Tipe_Kendaraan**
  - Pada fitur ini, terdapat 10 kategori dengan komposisi huruf alfabet A-E, masing-masing *lower* dan *upper case*
  - Terlihat juga bahwa distribusi kategori untuk yang *upper case* dan *lower case* saja cukup mirip
  - Diduga bahwa kategori huruf yang sama tetapi dengan *case* yang berbeda sesungguhnya merepresentasikan tipe kendaraan yang serupa
  - Dilakukan penggabungan kategori sehingga pada akhirnya hanya terdapat 5 kategori unik



Data Preprocessing: Imputasi *Missing Value*

- Imputasi
  - Dilakukan imputasi *median* untuk fitur **Jarak_Perjalanan**
  - Dilakukan imputasi *mean* untuk fitur numerik lainnya
  - Dilakukan imputasi *most frequent* untuk fitur kategorikal



- Label Encoding
  - Dilakukan *label encoding* pada fitur kategorikal agar dimensi variabel prediktor tidak terlalu banyak

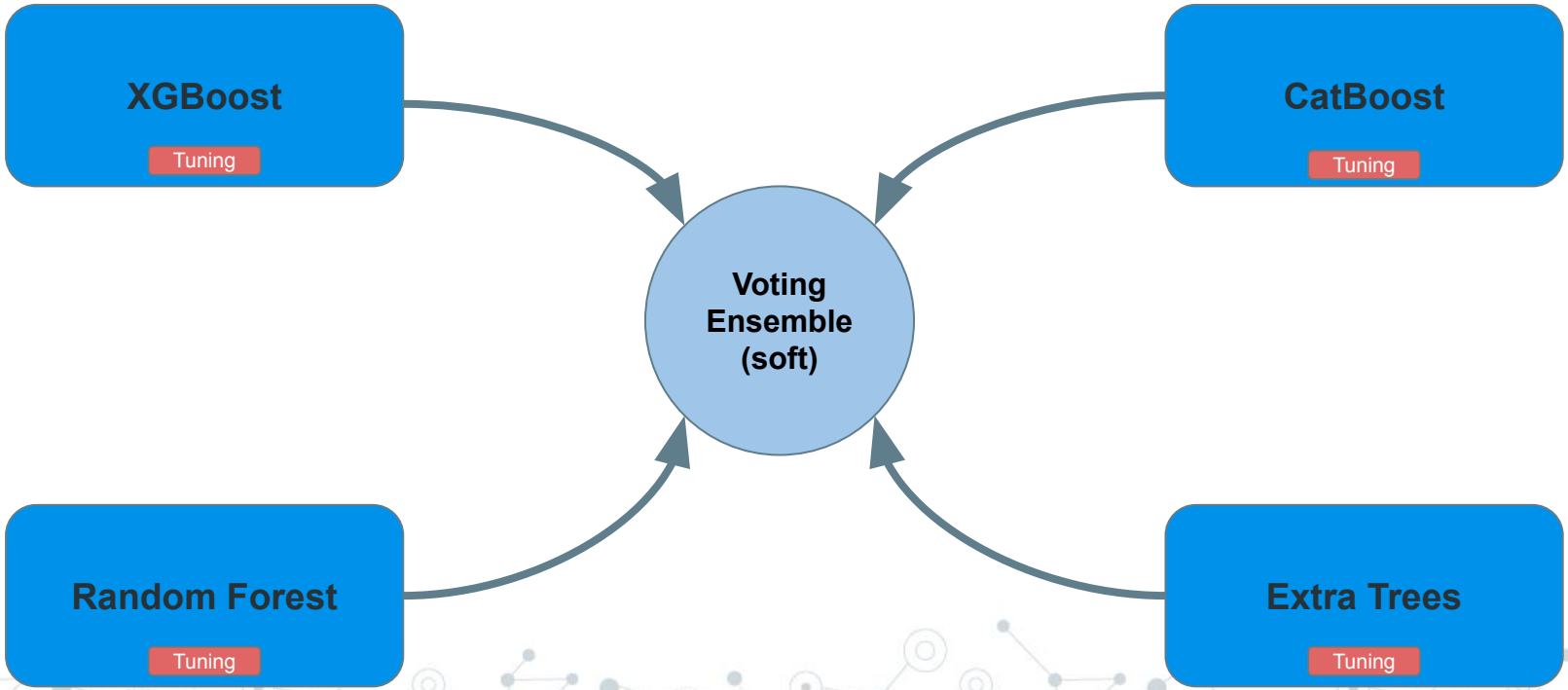
4.

Feature Engineering & Modelling

Feature Engineering

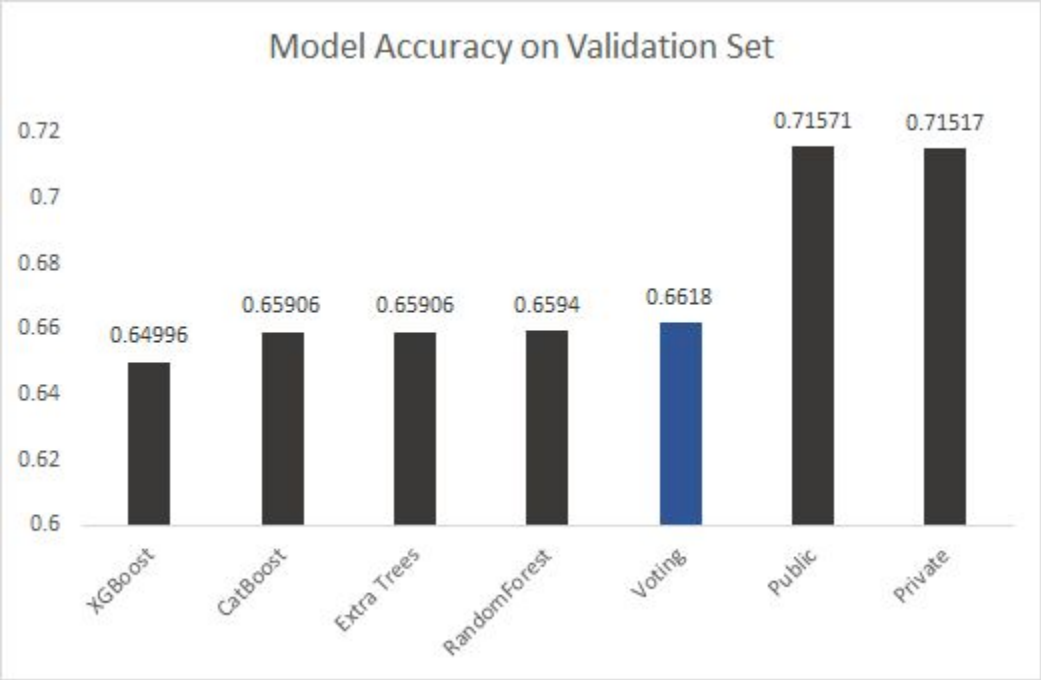
Rasio Jarak / Indeks	= Jarak_Perjalanan ÷ Indeks_Gaya_Hidup
Kategori Jarak	= membuat Jarak_Perjalanan menjadi kategorikal (binning)
NaN per Row	= jumlah *missing values* per baris
Has NaN	= indikator apakah terdapat *missing values* per baris

Modelling

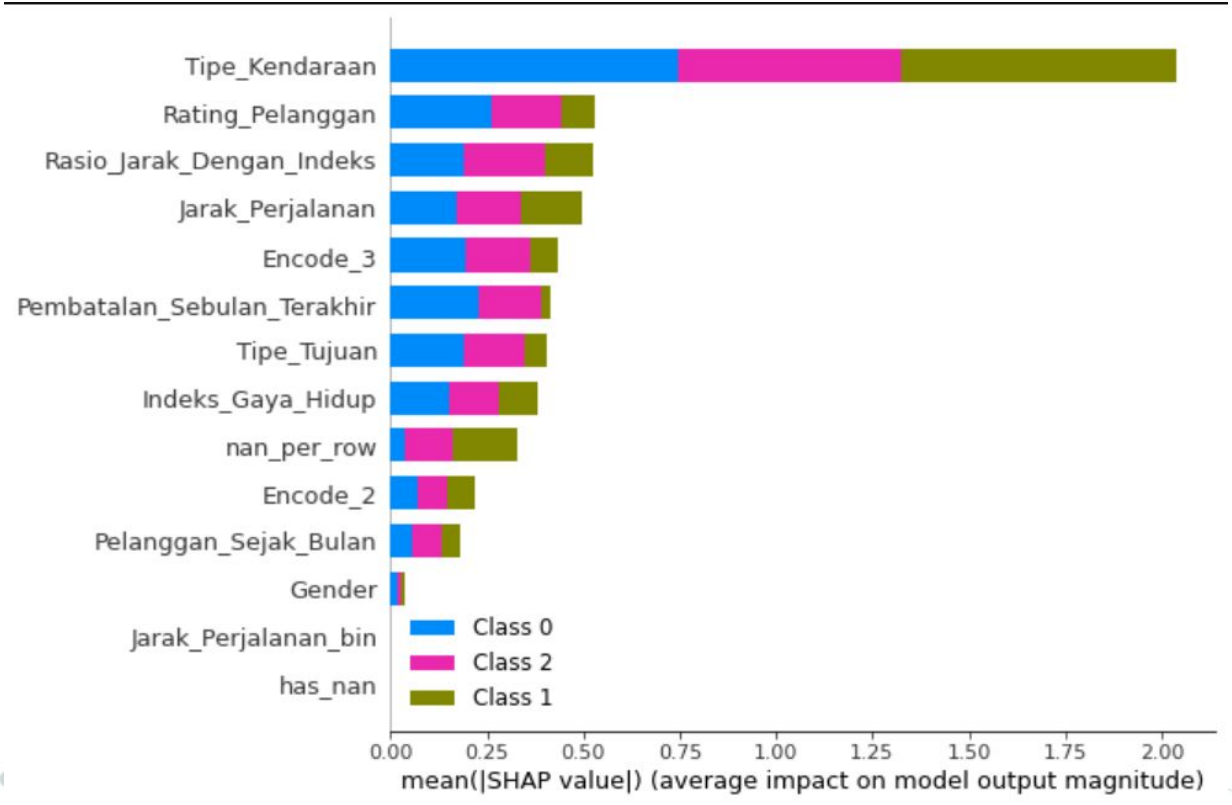

XGBoost
Tuning
CatBoost
Tuning
Voting Ensemble (soft)
Random Forest
Tuning
Extra Trees
Tuning

5.
Evaluation

Evaluation


Model Accuracy on Validation Set
0.72
0.7
0.68
0.66
0.64
0.62
0.6
0.64996
0.65906
0.65906
0.6594
0.6618
0.71571
0.71517
XGBoost
CatBoost
Extra Trees
RandomForest
Voting
Public
Private

Evaluation


Tipe_Kendaraan
Rating_Pelanggan
Rasio_Jarak_Dengan_Indeks
Jarak_Perjalanan
Encode_3
Pembatalan_Sebulan_Terakhir
Tipe_Tujuan
Indeks_Gaya_Hidup
nan_per_row
Encode_2
Pelanggan_Sejak_Bulan
Gender
Jarak_Perjalanan_bin
has_nan
Class 0
Class 2
Class 1
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
mean(|SHAP value|) (average impact on model output magnitude)

6. Conclusion

Model & Prediksi Memberi Indikasi

- Walau belum memiliki akurasi yang tinggi dengan prediksi yang dihasilkan dengan model terkait, terdapat indikasi bahwa peninjauan terhadap data *trip* ini benar-benar dapat menjadi acuan untuk mengelompokkan pelanggan sehingga pelayanan dari J Taxi dapat ditargetkan dan ditingkatkan kualitasnya
- *Future Studies* dapat difokuskan untuk mengembangkan model yang lebih baik untuk melakukan prediksi, ataupun mengolah data dalam aspek *preprocessing* atau *feature engineering* dengan lebih baik untuk memperoleh metode atau *pipeline* yang pada akhirnya dapat dijadikan elemen terintegrasi dalam menentukan *value delivery* dalam J Taxi

Thanks!

Any questions?