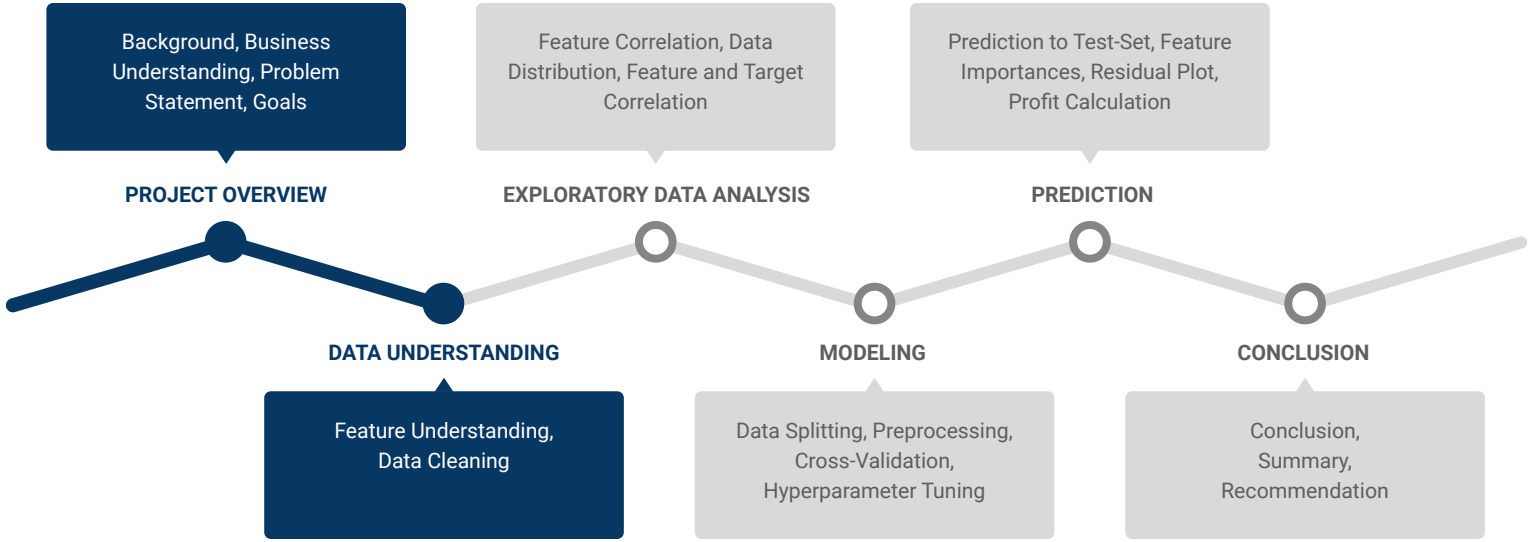
Bike Rental Prediction with Machine Learning for a Bike-Sharing Business

Capstone Project Module 3 - Osvaldo Sirait

OUTLINE


Background, Business Understanding, Problem Statement, Goals
PROJECT OVERVIEW
Feature Correlation, Data Distribution, Feature and Target Correlation
EXPLORATORY DATA ANALYSIS
Prediction to Test-Set, Feature Importances, Residual Plot, Profit Calculation
PREDICTION
DATA UNDERSTANDING
Feature Understanding, Data Cleaning
MODELING
Data Splitting, Preprocessing, Cross-Validation, Hyperparameter Tuning
CONCLUSION
Conclusion, Summary, Recommendation

Project Overview

BACKGROUND

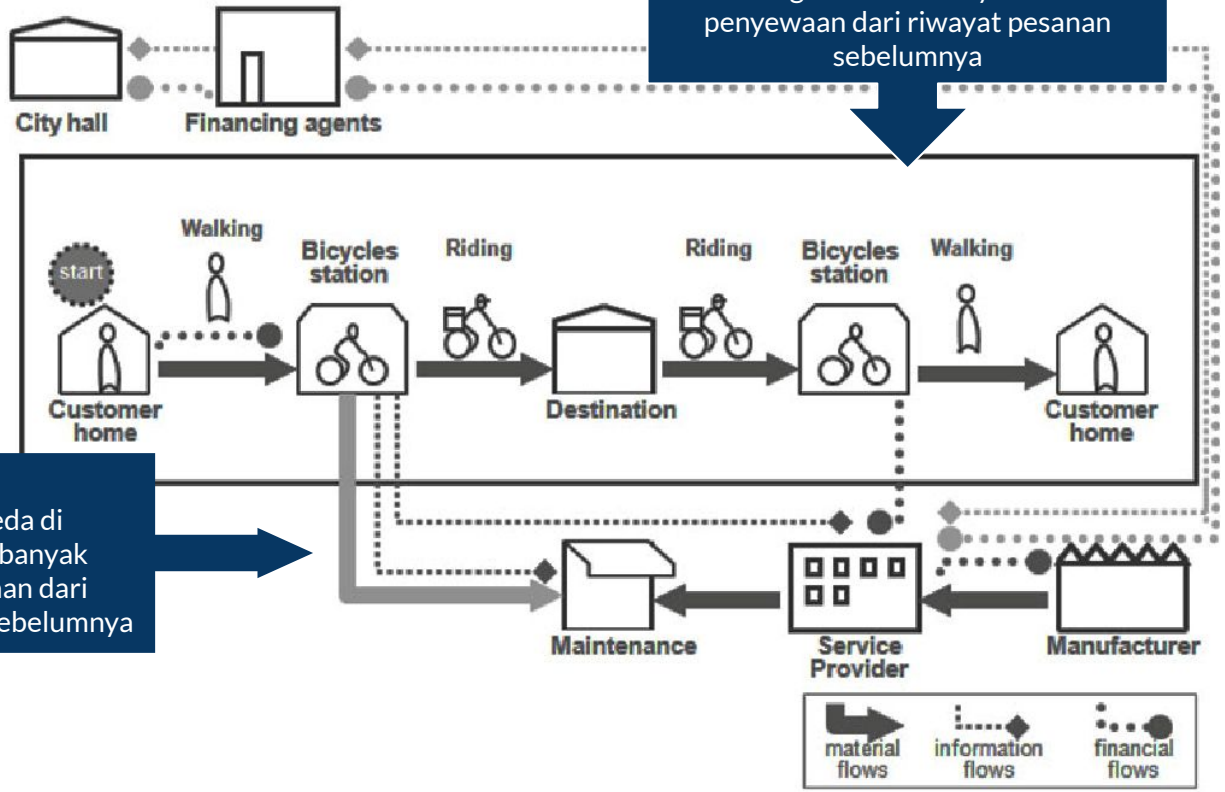
Bike-sharing system adalah sistem generasi terbaru dari bisnis persewaan sepeda tradisional di mana seluruh proses, mulai dari membership, penyewaan sepeda, dan pengembalian sepeda tersebut, telah menjadi otomatis. Melalui sistem ini, pengguna dapat dengan mudah menyewa sepeda dari posisi tertentu dan melakukan pengembalian di posisi lain.

Aspek business bike-sharing system:

1. **Enterprise**
   Pemilik Bisnis dan/atau Penyedia Layanan
2. **Infrastruktur**
   Docking Station / Tempat pengambilan & pengembalian sepeda
3. **Sepeda**
   Armada sepeda yang aman, tahan lama, dan cocok untuk jarak dekat
4. **Membership**
   Pay-per-ride, sewa harian, sewa bulanan, sewa tahunan
5. **Mobile Apps**
   Saran melakukan pemesanan, melihat riwayat pemesanan, dll
6. **Maintenance**
   Pemeliharaan rutin, perbaikan, pemantauan sepeda & docking station

BUSINESS MODEL

Penyedia layanan menyediakan sepeda di docking station sebanyak rata-rata penyewaan dari riwayat pesanan sebelumnya



Penyedia layanan menyediakan sepeda di docking station sebanyak rata-rata penyewaan dari riwayat pesanan sebelumnya

PROJECT GOALS

Business Problems

Meskipun ada perkembangan bisnis dan peningkatan ketertarikan yang cukup signifikan, **demand yang ada tidak selalu stabil.** Maka diperlukan **optimasi PPIC** agar bisa menjaga rasio investment yang lebih terukur sehingga **meminimalkan kerugian**.

Project Objectives

Data Understanding

ATTRIBUTES INFORMATION

- Dataset merupakan riwayat dari banyaknya sepeda yang dipinjam oleh pelanggan di tiap hari nya.
- Setiap baris data merepresentasikan informasi terkait kondisi lingkungan berdasarkan hari nya.
- Dataset has no missing value and no duplicates

Attribute	Data Type	Description
dteday	Object	Tahun-Bulan-Tanggal riwayat peminjaman sepeda
hum	float64	Nilai humiditas yang sudah dinormalisasikan (nilai dibagi 100 (max))
holiday	int64	Hari libur / Tidak
season	int64	Musim (1: winter, 2: spring, 3: summer, 4: fall)
atemp	float64	Indeks Kepanasan (Celcius) yang sudah dinormalisasikan. Nilai diperoleh melalui (t-tmin)/(tmax-tmin), tmin=-16, t_max=+50 (hanya dalam skala per jam)
temp	float64	Suhu (Celcius) yang sudah dinormalisasikan. Nilai diperoleh melalui (t-tmin)/(tmax-tmin), tmin=-8, t_max=+39 (hanya dalam skala per jam)
hr	int64	Jam (00:00 hingga 23:00)
casual	int64	Banyaknya peminjam sepeda yang tidak terdaftar sebagai member
registered	int64	Banyaknya peminjam sepeda yang terdaftar sebagai member
cnt	int64	Total banyaknya peminjam sepeda
weathersit	int64	Cuaca hari tersebut, dimana: 1 (Cerah, Sedikit awan, Sebagian berawan), 2 (Kabut + Mendung, Kabut + Awan Terpisah, Kabut + Sedikit awan, Berkabut), 3 (Salju Ringan, Hujan Ringan + Badai Petir + Awan Tersebar, Hujan Ringan + Awan Tersebar), 4 (Hujan Lebat + Palet Es + Badai Petir + Kabut, Salju + Kabut)

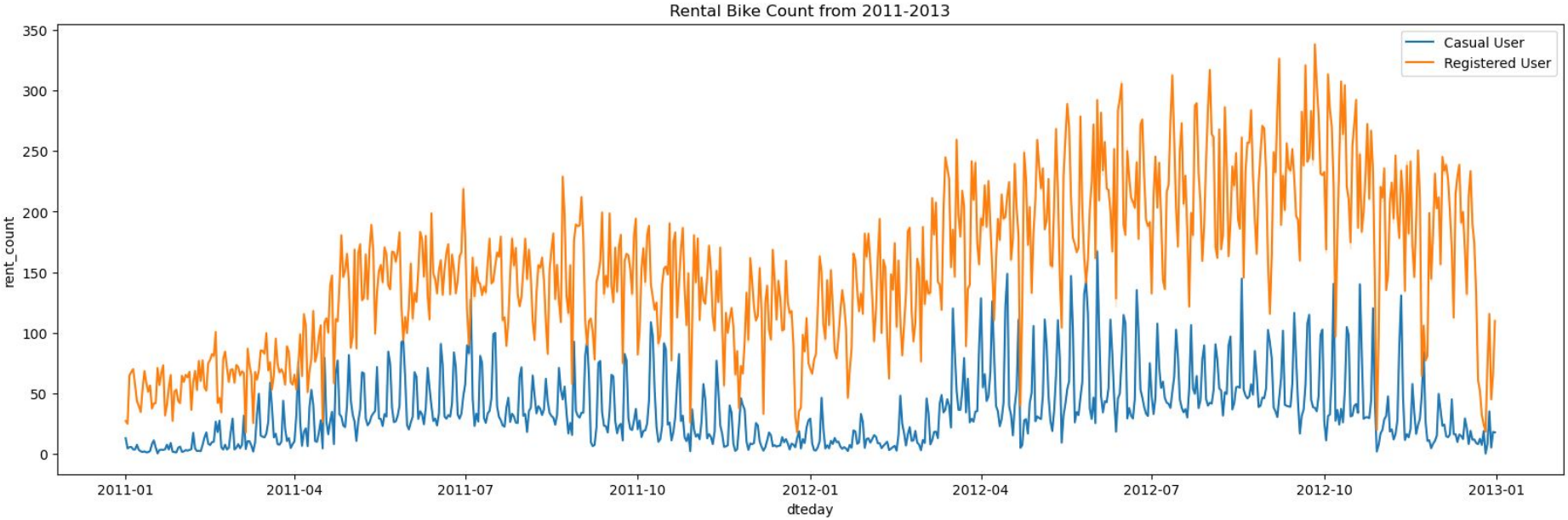
EVALUATION METRICS

Karena target yang diharapkan berupa jumlah sepeda, maka pada project ini digunakan analytical approach dengan model regresi. Untuk mengevaluasi kinerja Machine Learning yang dihasilkan, maka digunakan beberapa metrik yaitu RMSE, MAE, dan MAPE.

Priority	Evaluation Metrics	Description	Indicator
1	**MAE**	Mean Absolute Error digunakan untuk mengukur sejauh mana perbedaan antara nilai yang diprediksi oleh suatu model atau algoritma dengan nilai yang sebenarnya, tanpa memperhatikan arah perbedaan	Semakin kecil nilai MAE (mendekati 0), semakin bagus model nya
2	**RMSE**	Root Mean Square Error adalah akar rata-rata dari selisih kuadrat antara nilai yang diprediksi dan nilai yang sebenarnya. Metrik ini digunakan untuk mengukur sejauh mana perbedaan antara nilai yang diprediksi oleh suatu model statistik atau algoritma dengan nilai yang sebenarnya	Semakin kecil nilai RMSE (mendekati 0), semakin bagus model nya
3	**MAPE**	Mean Absolute Percentage Error digunakan untuk mengukur persentase kesalahan prediksi rata-rata dari suatu model atau algoritma. Metrik ini merupakan metrik yang paling mudah diinterpretasi dan dijelaskan ke masyarakat umum	Semakin kecil persentase kesalahan, semakin baik performa model dalam memprediksi nilai yang sebenarnya

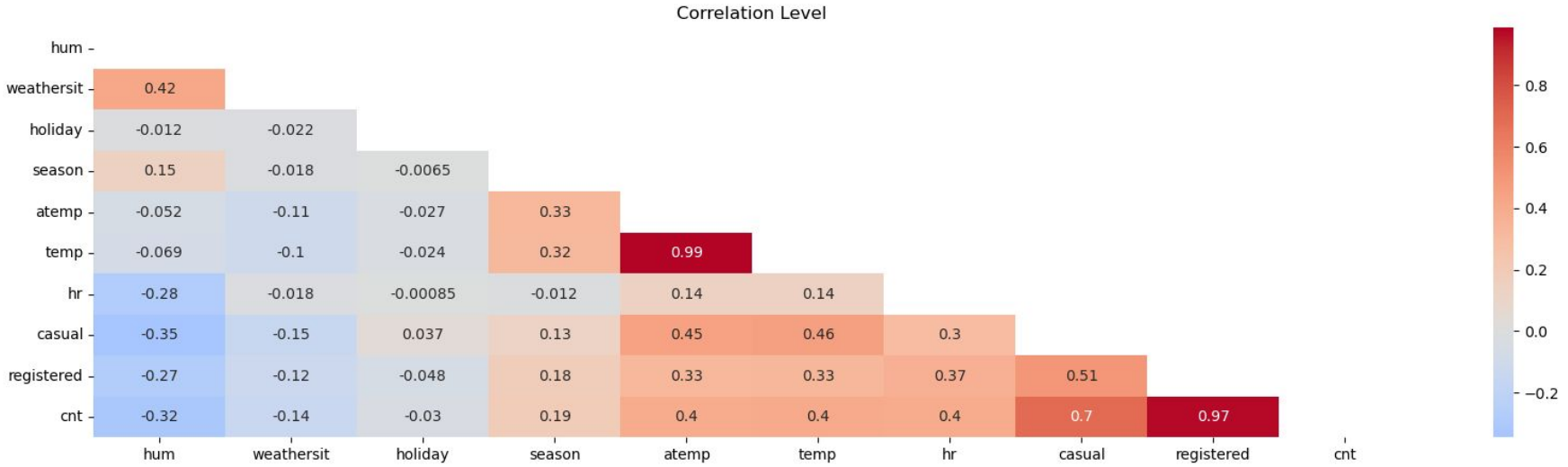
Exploratory Data Analysis

Bike-Rental Year-on-Year Trend



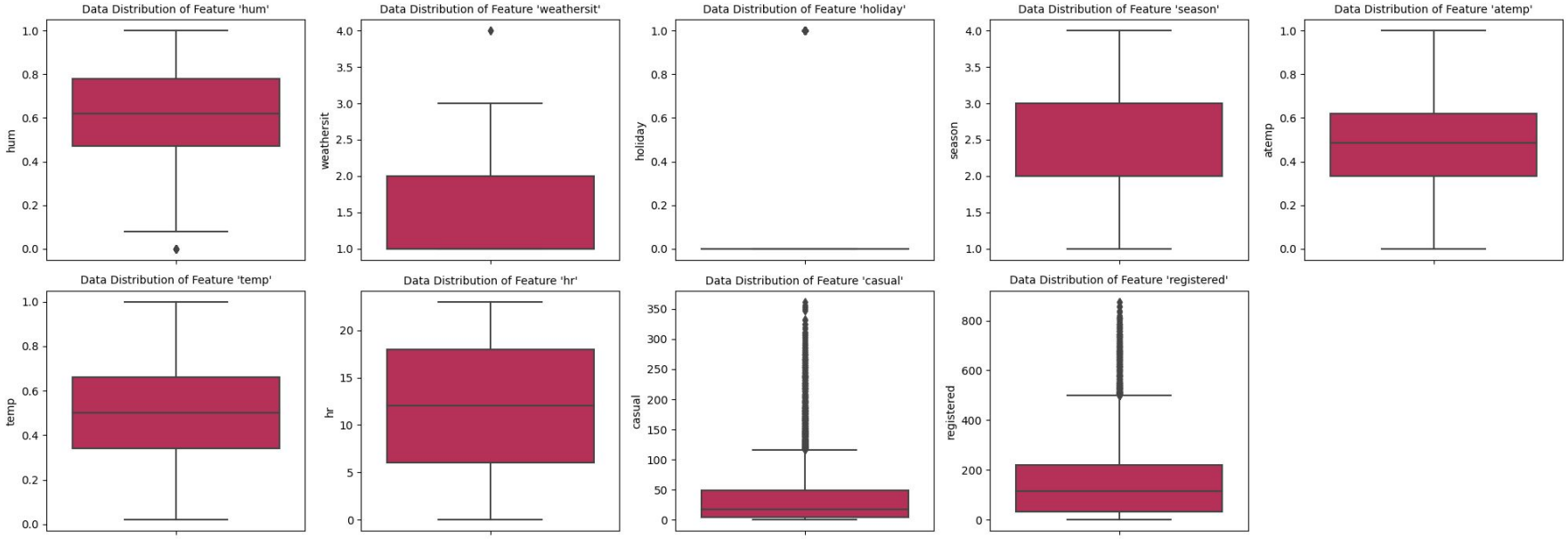
Demand terhadap sepeda sangat fluktuatif, baik untuk pelanggan yang terdaftar sebagai member maupun tidak. Akan tetapi, ada kecenderungan peningkatan jumlah peminjaman sepeda dari tahun 2011 ke 2013.

Feature Correlation



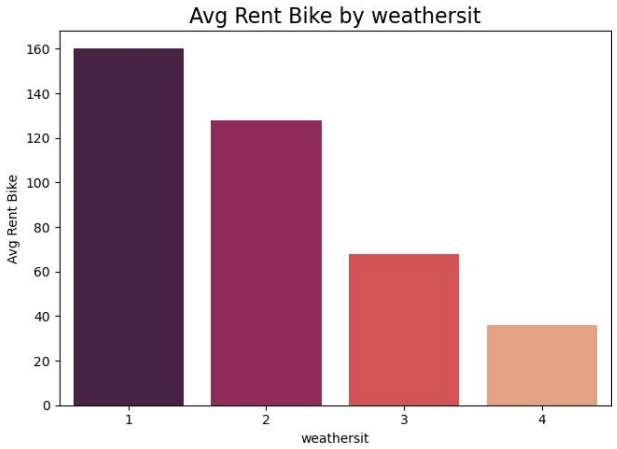
Ada hubungan positif dan kuat antara 'cnt' (total sepeda disewakan) dengan 'casual' dan 'registered', karena total sepeda yang disewa adalah hasil penjumlahan antara 'casual' dan 'registered'.

Data Distribution

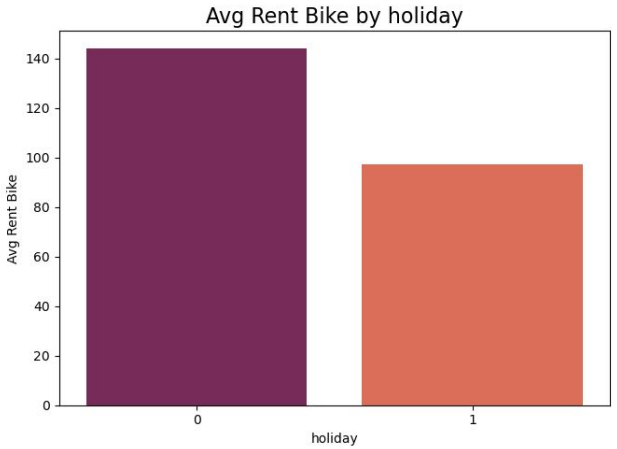


Persebaran data feature-feature yang terdapat di data set cenderung tidak normal. Akan tetapi, hanya kolom 'casual' dan 'registered' yang memiliki banyak outlier. Hal ini kemungkinan disebabkan karena (pada beberapa feature seperti 'temp' dan 'atemp'), data sudah di normalisasi. Sedangkan 'casual' dan 'registered' merupakan data actual.

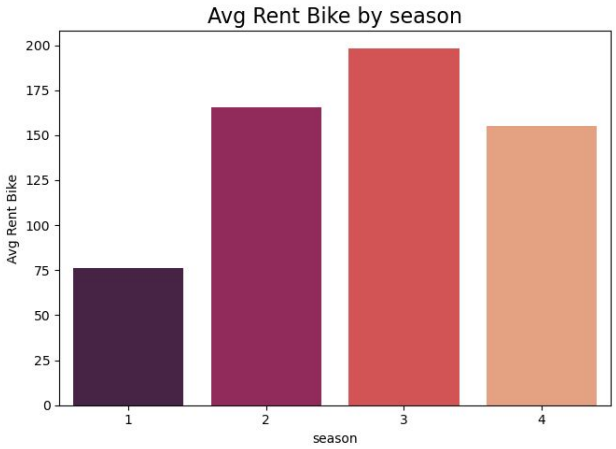
Bike-Rental per Feature



- Sepeda paling banyak disewa pada cuaca Cerah, Sedikit awan, Sebagian berawan
- Sepeda paling sedikit disewa ketika cuaca Hujan Lebat + Palet Es + Badai Petir + Kabut, Salju + Kabut
- Semakin buruk cuaca nya, semakin sedikit pelanggan yang sewa sepeda

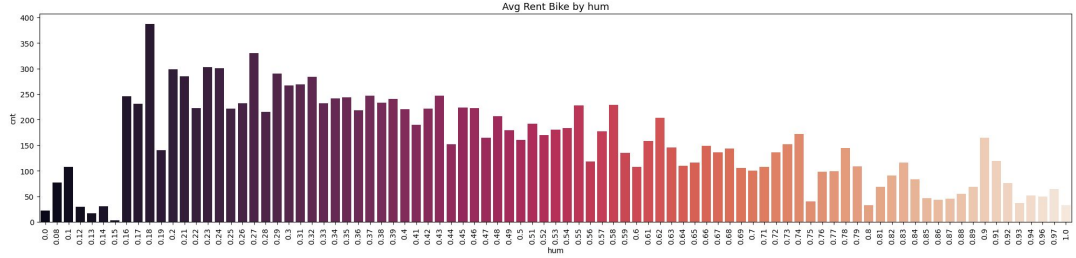


- Sepeda cenderung disewa pada hari yang bukan holiday / bukan hari libur



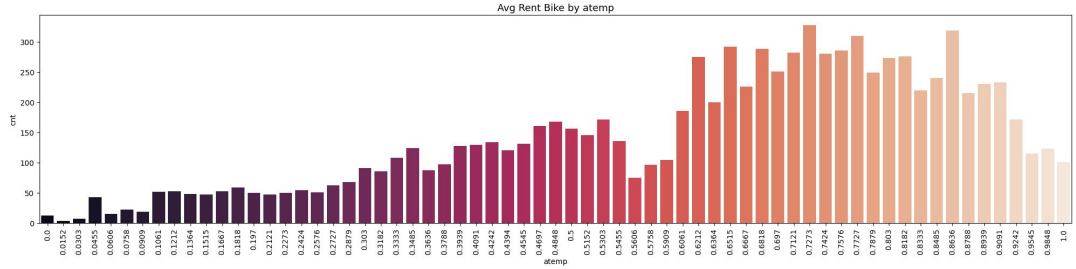
- Sepeda paling banyak disewa pada musim panas (summer)
- Sepeda paling sedikit disewa pada musim dingin (winter)

Bike-Rental per Feature



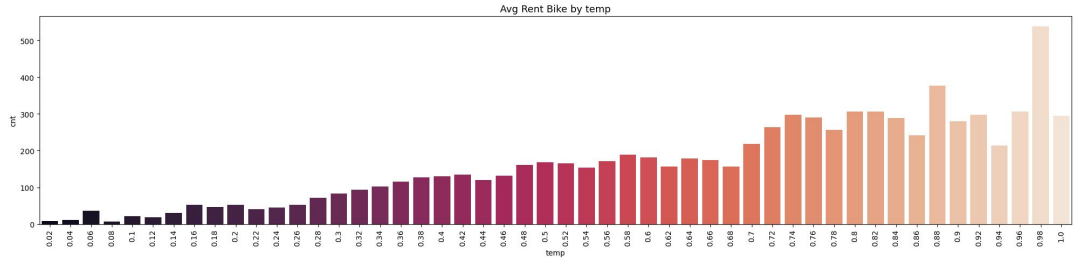
Berdasarkan Feature 'hum'

- Sepeda cenderung banyak disewa pada humidity rendah (0.16 - 0.32)
- Sepeda cenderung sedikit disewa pada humidity tinggi (diatas 0.70)



Berdasarkan Feature 'atemp'

- Sepeda cenderung banyak disewa pada index kepanasan tinggi (0.6 - 0.8)
- Sepeda cenderung sedikit disewa pada index kepanasan rendah (dibawh 0.5)
- Ada kecenderungan semakin tinggi index kepanasan, semakin banyak sepeda yang disewakan



Berdasarkan Feature 'temp'

- Sepeda cenderung banyak disewa pada suhu tinggi
- Sepeda cenderung sedikit disewa pada suhu rendah
- Ada kecenderungan semakin tinggi suhu, semakin banyak sepeda yang disewakan

Modeling

1) Defining Feature & Target

Features

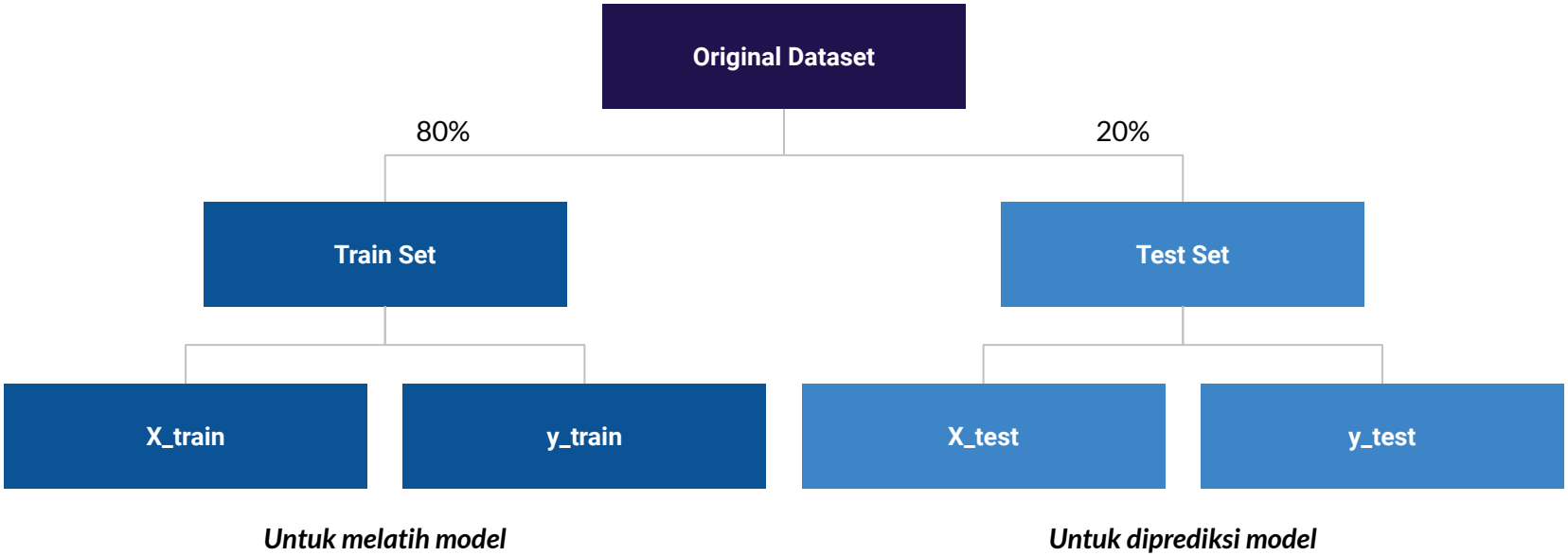
- year
- month
- day
- humidity
- weathersit
- holiday
- season
- atemp
- temp
- hour

Target

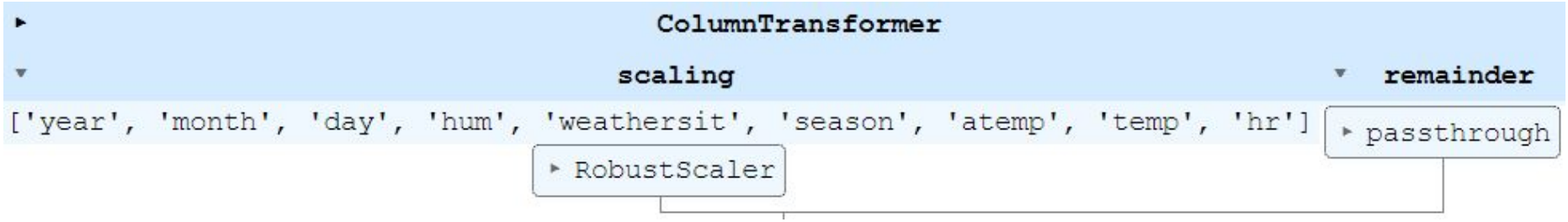
- cnt (count bike rent)

Remarks
Unused column from original dataset:
'dtday', 'casual', 'registered'

2) Data Splitting



3) Encoding & Normalization



Tidak ada encoding yang dilakukan karena tidak ada fitur pada dataset yang membutuhkan encoding.

Scaling sebelum hyperparameter tuning menggunakan RobustScaler: pengurangan nilai data dengan median, kemudian membaginya dengan interquartile range (IQR) - lebih stabil terhadap outlier.

4) Cross Validation

Beberapa model dilatih dengan menggunakan data latih dan kemudian dievaluasi dengan menggunakan data uji

Hasil Cross Validation digunakan untuk melihat seberapa baik model yang dibangun akan berkinerja pada data set baru.

Dataset awal dibagi menjadi 5 subset ("fold"). Pada setiap iterasi, setiap fold digunakan sebagai data uji, subset lainnya sebagai data latih

Single Model

- Linear Regression
- Lasso
- Ridge
- K-Nearest Neighbors
- Decision Tree

Ensemble Various

- Voting
- Stacking

Ensemble Similar

- Bagging (KNN)
- Random Forest
- Gradient Boosting
- Extreme Gradient Boosting
- Adaptive Boosting

*Semua model menggunakan parameter default

4) Cross Validation

	algo	mean_rmse	std_rmse	mean_mae	std_mae	mean_mape	std_mape
8	randomforest	-84.685610	2.440495	-54.971750	1.213284	-0.788928	0.070766
11	xgboost	-85.172804	2.447151	-56.778842	1.170770	-0.914524	0.060868
10	gboost	-87.924215	2.506002	-61.051955	1.369267	-1.026370	0.108991
6	stacking	-98.644428	3.069633	-65.628075	1.560265	-1.289009	0.150479
7	bagging	-114.237066	3.184636	-78.874658	1.976528	-1.894769	0.230475
4	tree	-114.614666	4.676735	-68.812721	1.987142	-0.882975	0.050046
9	adaboost	-115.083731	4.353021	-95.823236	5.931144	-3.457713	0.705338
3	knn	-115.793979	2.917295	-79.738667	1.867388	-1.906098	0.235405
5	voting	-116.038637	3.652898	-85.261895	2.130339	-2.403386	0.197058
2	ridge	-143.243665	4.924191	-107.292617	2.807782	-3.344205	0.252365
0	linreg	-143.244731	4.924120	-107.295618	2.808257	-3.344156	0.252298
1	lasso	-143.350895	5.031010	-107.055751	2.897482	-3.350911	0.254485

CV Score mendekati 0, STD rendah

5) Hyperparameter Tuning

proses mengoptimalkan nilai-nilai hyperparameter dalam sebuah model machine learning untuk mencapai performa yang lebih baik.

Gradient Boosting	Extreme Gradient Boosting	Random Forest
5x RandomSearch tuning	***5x RandomSearch tuning***	***5x RandomSearch tuning***

	rmse	mae	mape
Tuning 5	-85.577508	-54.987786	-0.790025
Tuning 3	-84.328609	-55.363812	-0.846922
Tuning 2	-84.883867	-55.435969	-0.872686
Tuning 1	-84.972749	-55.453777	-0.874301
Tuning 4	-84.246712	-55.555064	-0.842782
Tuning 0	-87.924215	-61.051955	-1.026370

	rmse	mae	mape
Tuning 4	-84.579900	-54.245709	-0.789731
Tuning 2	-83.506126	-55.593890	-0.870779
Tuning 1	-83.516893	-55.679620	-0.863909
Tuning 3	-83.376026	-55.691577	-0.859254
Tuning 0	-85.172804	-56.778842	-0.914524
Tuning 5	-88.887583	-62.393202	-1.134333

	rmse	mae	mape
Tuning 5	-84.393240	-54.791987	-0.786013
Tuning 0	-84.685610	-54.971750	-0.788928
Tuning 4	-88.108895	-56.593319	-0.792918
Tuning 2	-86.887658	-58.965771	-0.926859
Tuning 3	-88.628856	-60.530571	-0.971314
Tuning 1	-104.016297	-73.168449	-1.467455

hyper-params:	n_estimators, min_samples_split, min_samples_leaf, max_features, max_depth, learning_rate, scaler	n_estimators,learning_rate, reg_alpha, subsample, colsample_bytree, max_depth, eval_metric, gamma, scaler	n_estimators, max_depth, n_samples, max_features, min_samples, max_leaf, bootstrap, scaler

5) Best Model

XGBoost (Extreme Gradient Boosting) adalah sebuah algoritma ensemble learning dengan konsep menggabungkan beberapa model prediktif yang buruk menjadi model yang lebih baik. Setiap model berikutnya dibangun untuk memperbaiki kesalahan prediksi model sebelumnya. Dalam hal ini, model-model yang dihasilkan disebut "weak learner" atau "base learner".

- Bekerja sequential/berurutan
- Menerapkan penanganan missing values dan regularisasi
- Menghasilkan model kompleks dan akurat

5) Best Model

Best Hyperparameter after Tuning for XGBoost Model

Hyperparameter	Description	Values
n_estimators	Jumlah pohon yang akan dibangun dalam model	662
subsample	Penentuan fraksi dari dataset yang akan digunakan dalam setiap iterasi	0.9 (90% dari dataset akan digunakan dalam setiap iterasi)
reg_alpha	Kekuatan regularisasi L1 pada fungsi objektif (penalti pada fitur)	1.0
max_depth	Penentuan kedalaman maksimum dari setiap pohon dalam model.	6
learning_rate	Penentuan seberapa cepat model belajar dari kesalahan prediksi pada setiap iterasi	0.03
gamma	Ambang batas yang digunakan untuk melakukan pruning pada pohon	0.5
eval_metric	Penentuan metrik evaluasi yang akan digunakan untuk mengukur kinerja model	'mlogloss' (multi-class logloss)
colsample_bytree	Penentuan fraksi dari fitur yang akan digunakan dalam setiap pohon	0.7 (70% fitur akan digunakan dalam setiap pohon)
random_state	Parameter ini digunakan untuk mengontrol inisialisasi generator angka acak dalam XGBoost	42

Prediction

Test Set Prediction

Gradient Boosting

	before_tuning	after_tuning	tuning_change
RMSE	86.835	79.344	-7.491
MAE	59.433	51.268	-8.165
MAPE	0.979	0.716	-0.263

Extreme Gradient Boosting

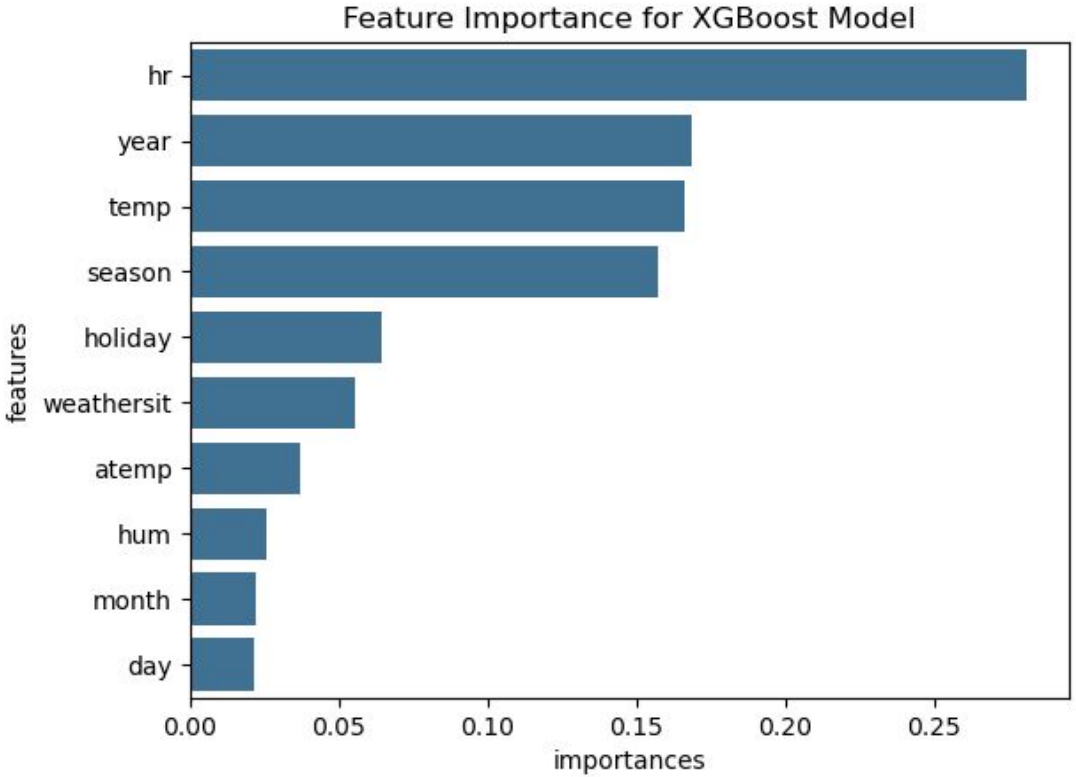
	before_tuning	after_tuning	tuning_change
RMSE	81.329	77.931	-3.398
MAE	54.254	50.559	-3.695
MAPE	0.841	0.716	-0.125

Random Forest

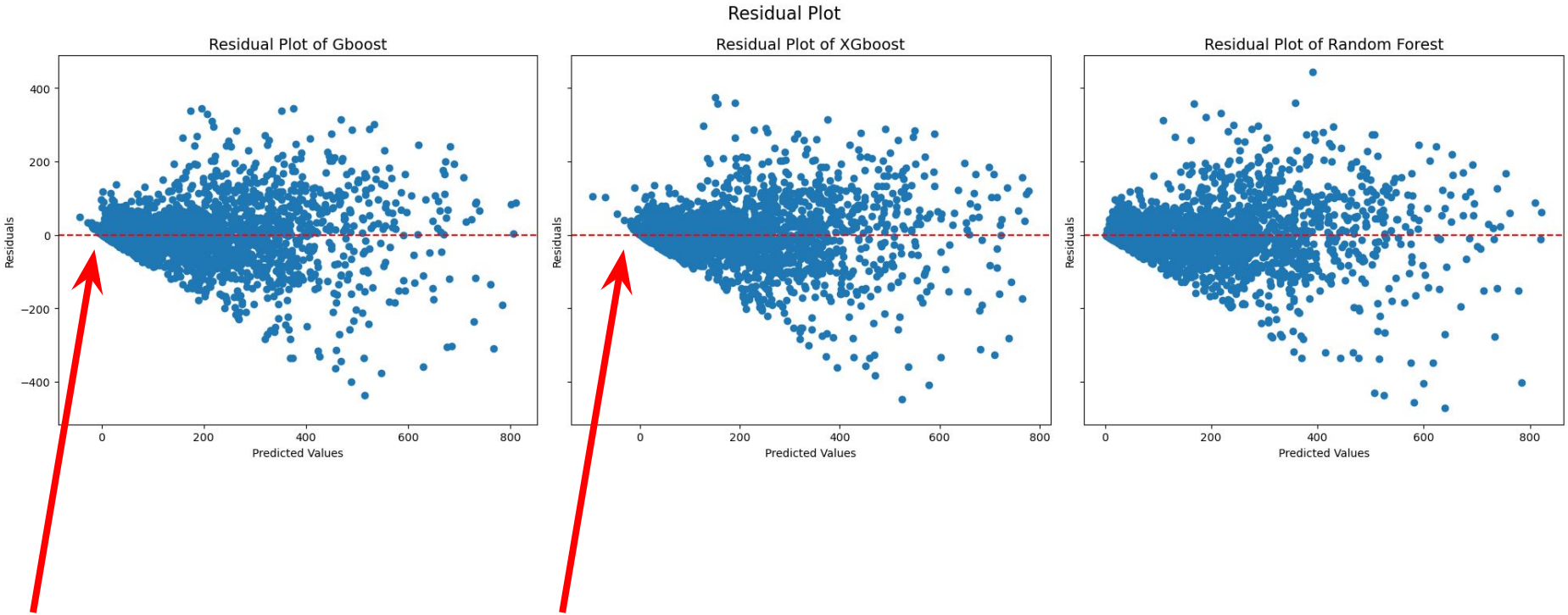
	before_tuning	after_tuning	tuning_change
RMSE	80.831	80.548	-0.283
MAE	52.415	52.341	-0.074
MAPE	0.725	0.720	-0.005

Paling rendah dibandingkan model lainnya

Feature Importances (XGBoost)



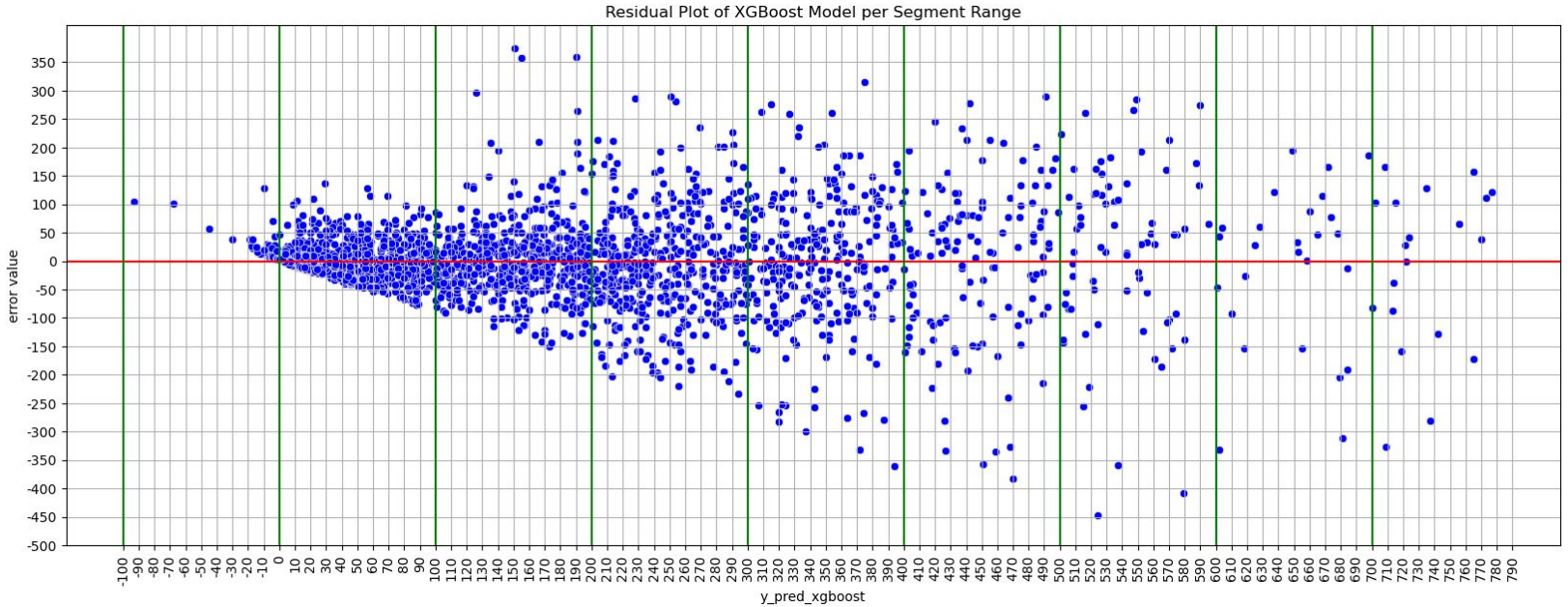
Residual Plot



Hasil prediksi lebih kecil dari y_actual

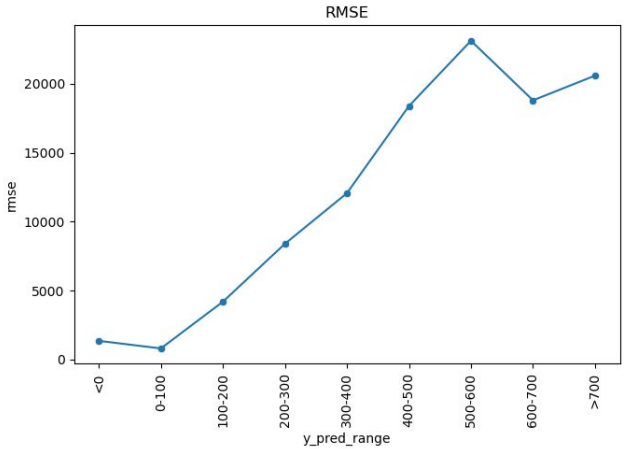
Hasil prediksi lebih kecil dari y_actual

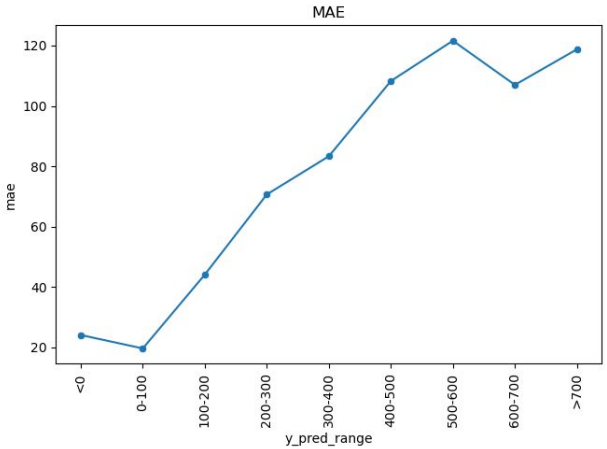
Residual Plot

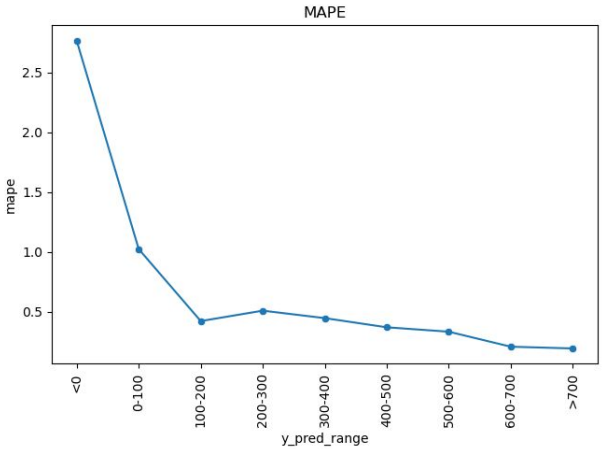


Pada rentang prediksi 200 hingga 400, hasil prediksi mulai memiliki error yang besar dan bervariasi.
Saat rentang prediksi diatas 400, error yang dihasilkan sangat besar dan sangat bervariasi.

Residual Plot







RMSE dan MAE berdasarkan range hasil prediksi memiliki nilai paling tinggi pada rentang 500-600, dimana terdapat pola semakin besar hasil prediksinya, semakin tinggi RMSE dan MAE nya hingga rentang 500-600. Nilai RMSE dan MAE turun pada range 600-700, tetapi naik kembali pada range diatas 700.

Profit Calculation

Business Case Assumption

- Harga sewa sebuah sepeda per customer baik terdaftar atau tidak = 5 dollar
- Biaya persiapan dan pengiriman sepeda ke titik penyewaan apabila banyak sepeda ditentukan 3 hari sebelumnya = 3 dollar
- Biaya persiapan dan pengiriman sepeda ke titik penyewaan apabila banyak sepeda ditentukan pada hari yang sama = 6 dollar
- Sebelum melakukan machine learning, perusahaan menyediakan sepeda per hari sebanyak rata-rata penyewaan sepeda sebelumnya. Apabila kekurangan, maka perusahaan akan menyediakan sepeda pada hari yang sama
- Dengan menggunakan machine learning, apabila ada prediksi banyak nya sepeda yang disewakan lebih sedikit daripada demand sebenarnya, maka diasumsikan bahwa pelanggan bersedia menunggu kedatangan sepeda ke docking station pada hari yang sama (sehingga tidak terjadi kehilangan sales/pelanggan tidak cancel order)

Profit Calculation

	Profit/Loss	increment (x)
XGBoost Model	520911	6.886
GBoost Model	515574	6.826
Random Forest Model	507885	6.739
No Machine Learning	-88497	0.000

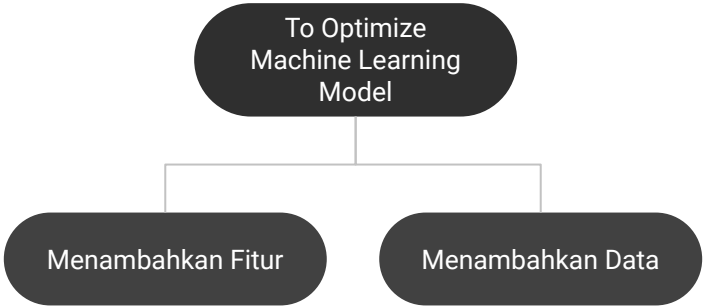
Dengan menggunakan Machine Learning XGBoost model, penyedia layanan dapat memperoleh keuntungan sebanyak 520,911 dollar, yaitu 6,89 kali lipat dari bisnis tanpa machine learning (profit meningkat sebanyak 688%)

Summary

Conclusion

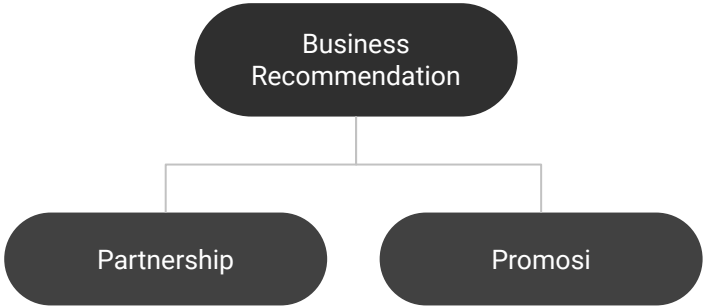
Model machine learning terbaik untuk prediksi jumlah sepeda yang disewakan	Extreme Gradient Boost
Fitur yang mempengaruhi machine learning model	Jam, tahun, suhu, dan musim
Metrik evalusai yang digunakan	MAE (prioritas), RMSE, dan MAPE
Akurasi prediksi	- Hasil prediksi dibawah 200 cukup baik - Hasil prediksi diantara 200-400 mulai memiliki error bervariasi - Hasil prediksi diatas 400 memiliki error besar dan bervariasi
Model machine learning dapat meningkatkan profit	- Tanpa machine learning kerugian sebesar 88,497 dollar - Machine learning XGBoost model keuntungan sebanyak 520,911 dollar

Recommendation



Dengan menambahkan fitur yang relevan, model dapat memiliki akses ke informasi tambahan yang dapat membantu dalam membuat prediksi yang lebih akurat.

Model masih memiliki performa yang buruk untuk memprediksi hasil yang cukup besar (diatas range 400). Dengan penambahan data yang hasil / target nya lebih besar, maka ada kesempatan untuk model machine learning ini menghasilkan performa yang lebih baik.



Hasil prediksi model ini dapat digunakan untuk memprediksi mobilitas sebuah kota, cocok untuk bekerja sama dengan pemerintah agar bisa mengatur lalu lintas kota dengan persiapan lebih baik

Pada hari-hari dimana machine learning memprediksi bahwa sepeda yang disewakan berjumlah sedikit, penyedia layanan dapat menjalankan promosi guna meningkatkan ketertarikan pelanggan untuk menyewa sepeda.

Thank You