

Bike Sharing Prediction



PREPARED BY HARYO DEWANTORO



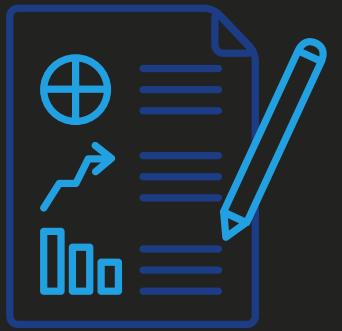
[Link GitHub](#)

1



Business Problem
Understanding

2



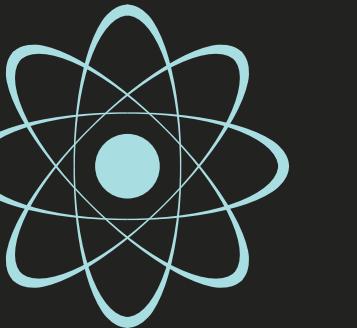
Data Understanding

3



Data Preprocessing

4



Modelling

5



Conclusion

6



Recommendation



Business Problem Understanding



● CONTEXT

Terdapat sekitar lebih dari 500 program bike-sharing di seluruh dunia dengan lebih dari 500.000 unit sepeda.

● PROBLEM STATEMENT

- Menyediakan jumlah unit sepeda yang cukup di setiap kondisi dan situasi.
- Jika jumlah unit tidak tercukupi, maka akan berdampak kepada hilangnya kepercayaan atau menurunnya jumlah pelanggan.
- Kehilangan pelanggan = menurunnya profit.

● GOAL

Membuat suatu model Machine Learning yang mampu memprediksi jumlah unit sepeda yang perlu tersedia di setiap kondisi dan situasi.

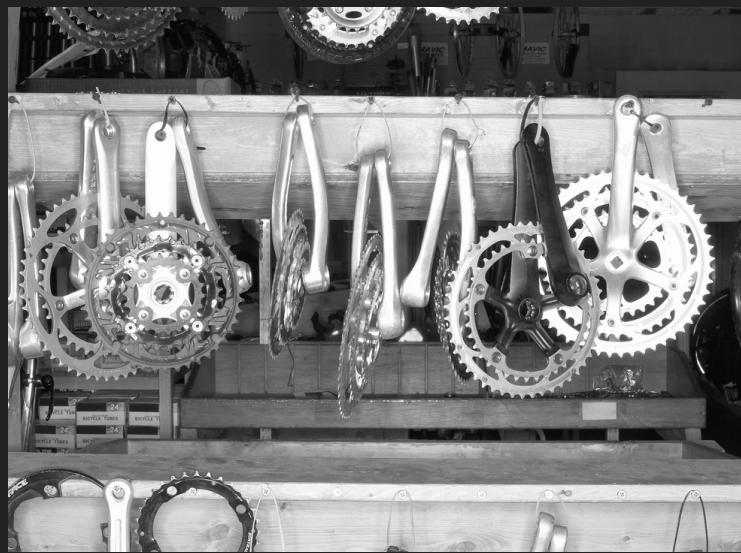
● ANALYTICS APPROACH

Melakukan analisa terhadap data untuk dapat menemukan pola pada fitur-fitur yang ada, yang membedakan satu kondisi dengan yang lainnya, serta bagaimana tiap kondisi tersebut mempengaruhi jumlah unit sepeda yang perlu tersedia. Selanjutnya, akan dibuat suatu model regresi yang bertujuan untuk menentukan jumlah unit sepeda yang perlu disediakan oleh Capital Bikeshare.

EVALUATION METRICS



Alasan Menggunakan MAE



MEAN ABSOLUTE ERROR (MAE)

MAE digunakan dalam project ini guna menghitung rataan nilai absolut dari error yang dihasilkan oleh model.

- MAE MEMILIKI INTERPRETASI YANG LEBIH MUDAH DIPAHAMI. NILAI MAE MEREPRESENTASIKAN RATA-RATA ABSOLUT DARI SELISIH ANTARA NILAI PREDIKSI DENGAN NILAI SEBENARNYA.
- MAE ADALAH METRIK YANG LEBIH TOLERAN TERHADAP PENCILAN (OUTLIERS) DIBANDINGKAN METRIK LAIN SEPERTI MEAN SQUARED ERROR (MSE)
- MAE BERSIFAT LINIER. INI BERARTI BAHWA SETIAP KESALAHAN PREDIKSI MEMILIKI KONTRIBUSI YANG PROPORSIONAL TERHADAP NILAI MAE.

DATA UNDERSTANDING



INFORMASI DATASET "BIKE SHARING"

Attribute	Data Type	Description
dteday	Object	Date
hum	Float	Normalized humidity (the values are divided to 100)
weathersit	Integer	1: Clear, Few clouds, Partly cloudy, Partly cloudy 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
holiday	Integer	0: Not holiday 1: Holiday
season	Integer	1: Winter 2: Spring 3: Summer 4: Fall
atemp	Float	"Feels like" temperature in Celsius
temp	Float	Normalized temperature in Celsius
hr	Integer	Hour (0 to 23)
casual	Integer	Count of casual users
registered	Integer	Count of registered users
cnt	Integer	Count of total rental bikes including both casual and registered users

DATA PREPROCESSING

PENYESUAIAN
NAMA KOLOM
DAN VALUE

MENGUBAH TIPE
DATA DAN
MEMISAHKAN
DATA 'DATE'

CHECKING
MISSING VALUE
AND DATA
DUPLIKAT

DROP
COLUMN(FEATUR
E SELECTION)

CHECKING DATA
CORRELATION
(MATRIX DAN VIF
SCORE)

CHECKING
OUTLIERS

CLEAN DATASET

MODELING

ENCODING

ONE HOT ENCODER:

1. WEATER
2. SEASON
3. YEAR

BINARY ENCODER:

1. DAY

CHOOSING BENCHMARK MODEL

model	MAE	MAPE	R2
Linear Regression	108.221783	1.413258	0.178131
KNN Regressor	45.703390	0.444276	0.833066
DecisionTree Regressor	42.734993	0.466782	0.833400
RandomForest Regressor	32.757434	0.329204	0.907222
Gradient Boosting	55.558862	0.474801	0.763944
XGBoost Regressor	28.881241	0.271619	0.931213

TRAIN AND TEST SPLITTING
70 TRAIN SET : 30 TEST SET



EXTREME GRADIENT BOOSTING

Extreme Gradient Boosting (xGBoost) adalah sebuah metode ensemble yang bekerja dengan membangun serangkaian model Decision Trees secara sekuensial.

Fungsinya adalah untuk meminimalkan kesalahan prediksi dan menghasilkan model yang dapat digunakan untuk memprediksi dengan akurasi yang tinggi.

PREDICT TO TEST SET USING BENCHMARK MODEL (PARAMETER BY DEFAULT)



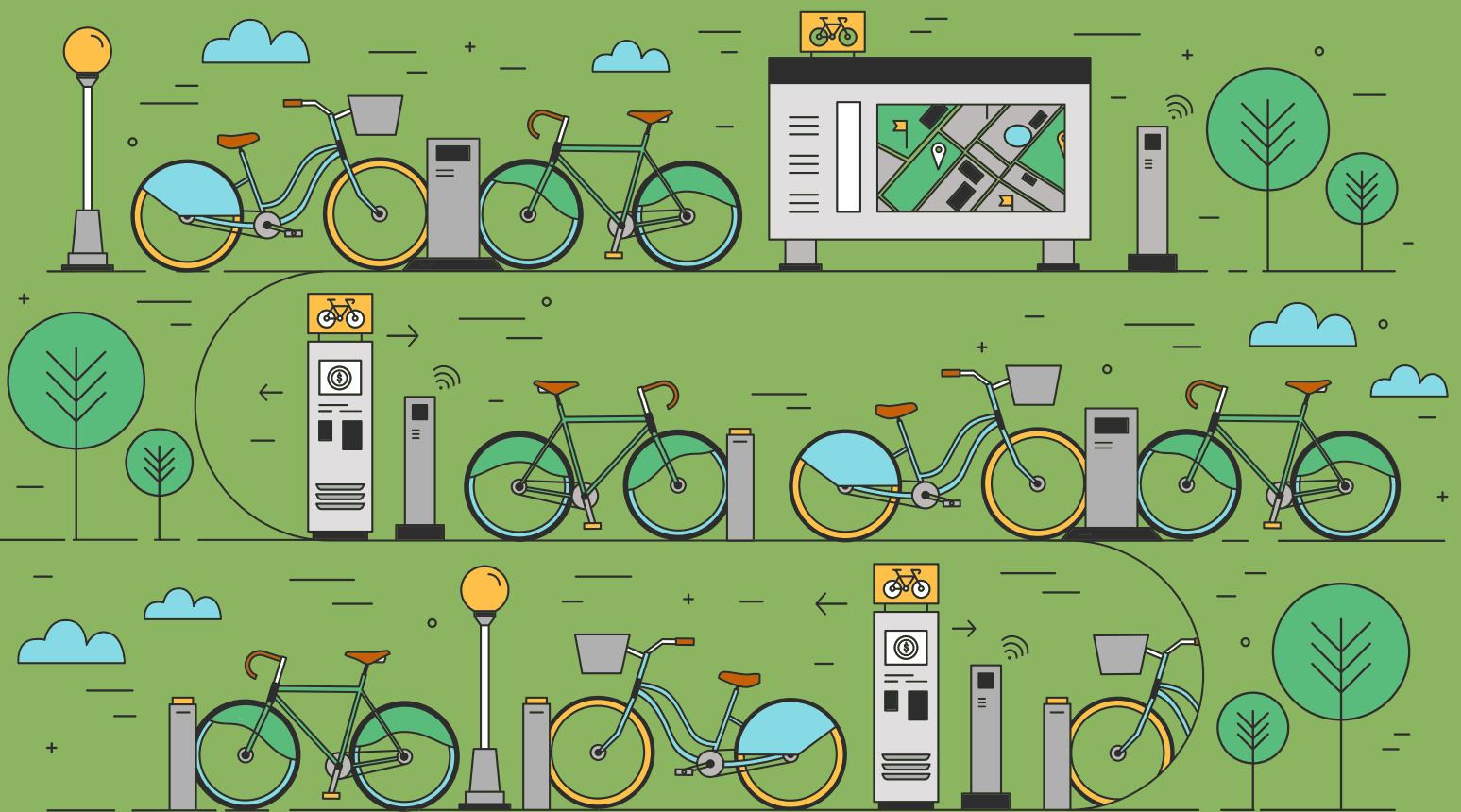
	MAE	MAPE	R-squared
XGB	27.079164	0.262935	0.94301

Terlihat bahwa ketika melakukan prediksi pada Test Set, XGBoost tetap memiliki performa yang baik. Dimana nilai MAE dan MAPE menurun (27.07), sedangkan nilai R-squared mengalami peningkatan (0.94)

HYPERPARAMETER TUNING (GRID SEARCH)



Pada tahap hyperparameter tuning ini digunakan metode Grid Search untuk mencoba seluruh kombinasi hyperparameter. Parameter yang digunakan akan dibatasi berupa `max_depth` (kedalaman pohon), `learning_rate` (ukuran step pada tiap iterasi), dan `n_estimators` (berapa kali iterasi). Hal ini dilakukan guna mencegah terjadinya overfitting yang bisa disebabkan jika model terlalu kompleks dan juga untuk menjaga efisiensi kinerja dari device (laptop) yang digunakan.



Parameter	Best Value
max_depth	8
learning_rate	0.1
n_estimators	200

	MAE	MAPE	R-squared
XGB	25.401036	0.250356	0.948515

HYPERPARAMETER TUNING RESULT

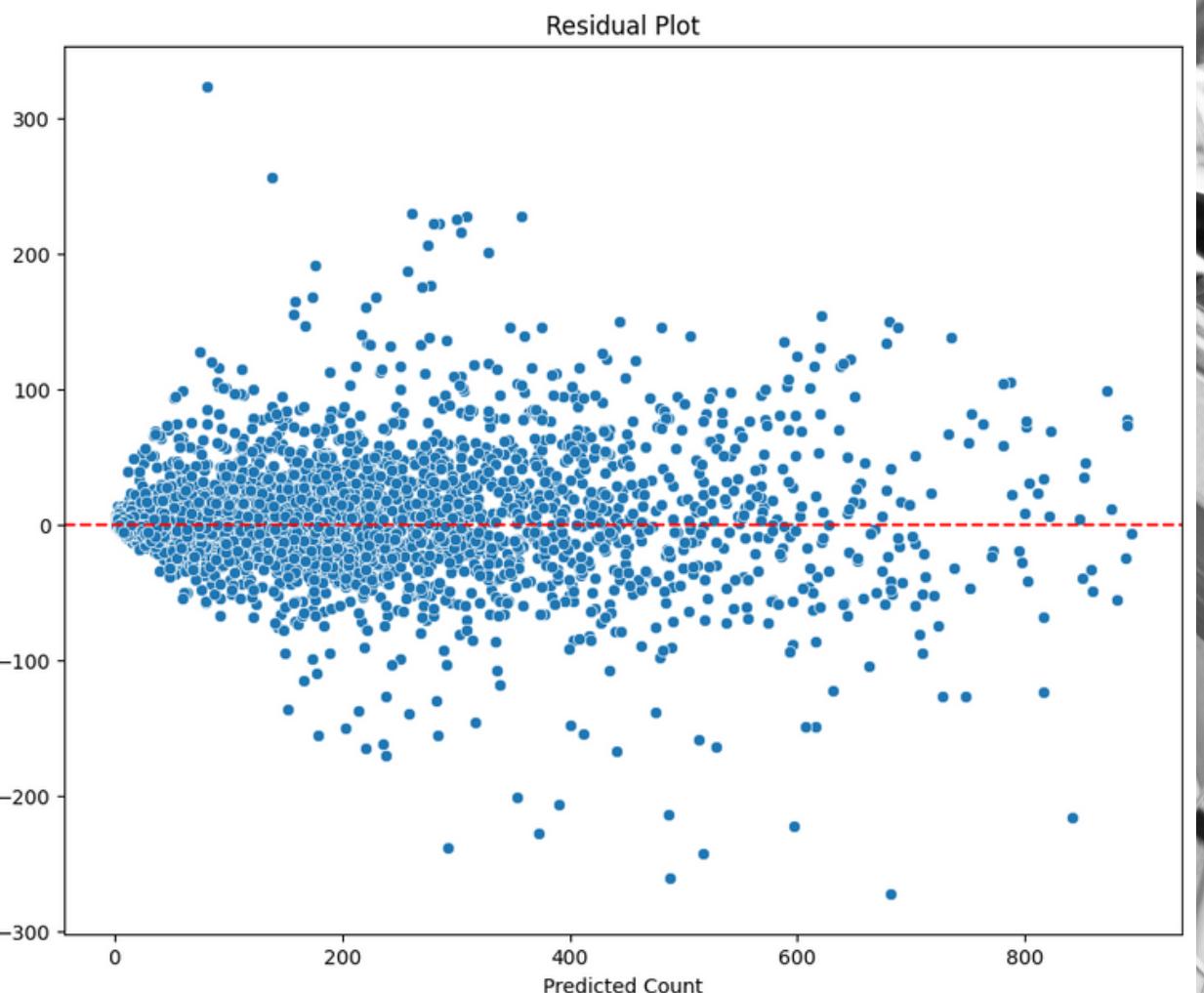
PREDICT TO TEST SET USING BEST PARAMETERS

PERFORMANCE COMPARISON

Berikut adalah komparasi performa model XGBoost sebelum dan sesudah dilakukan Hyperparameter Tuning:

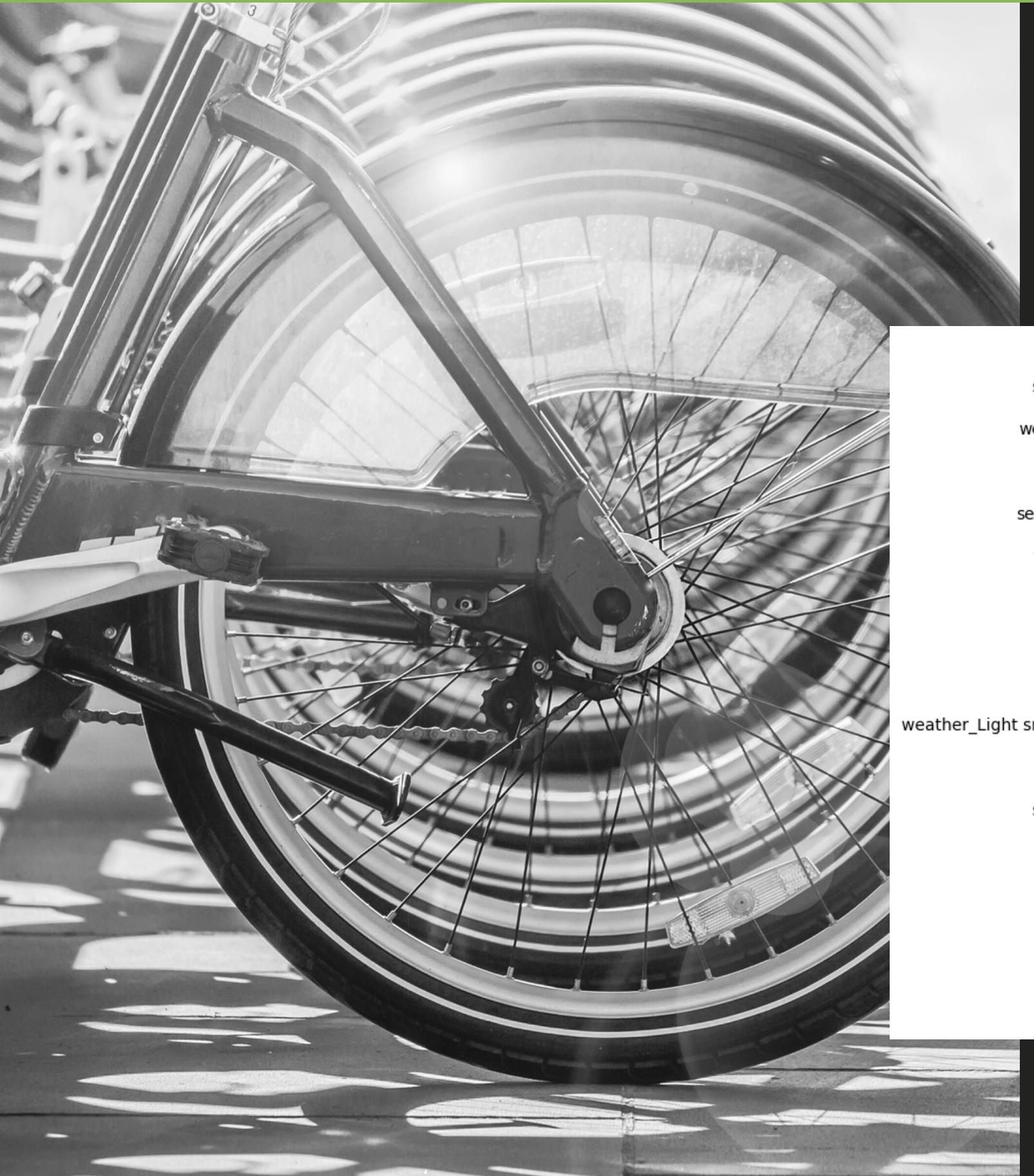
Model	MAE	MAPE	R-squared
XGBoost Test Before Tuning	27.079164	0.262935	0.94301
XGBoost Test After Tuning	25.401036	0.250356	0.948515

MODEL LIMITATION

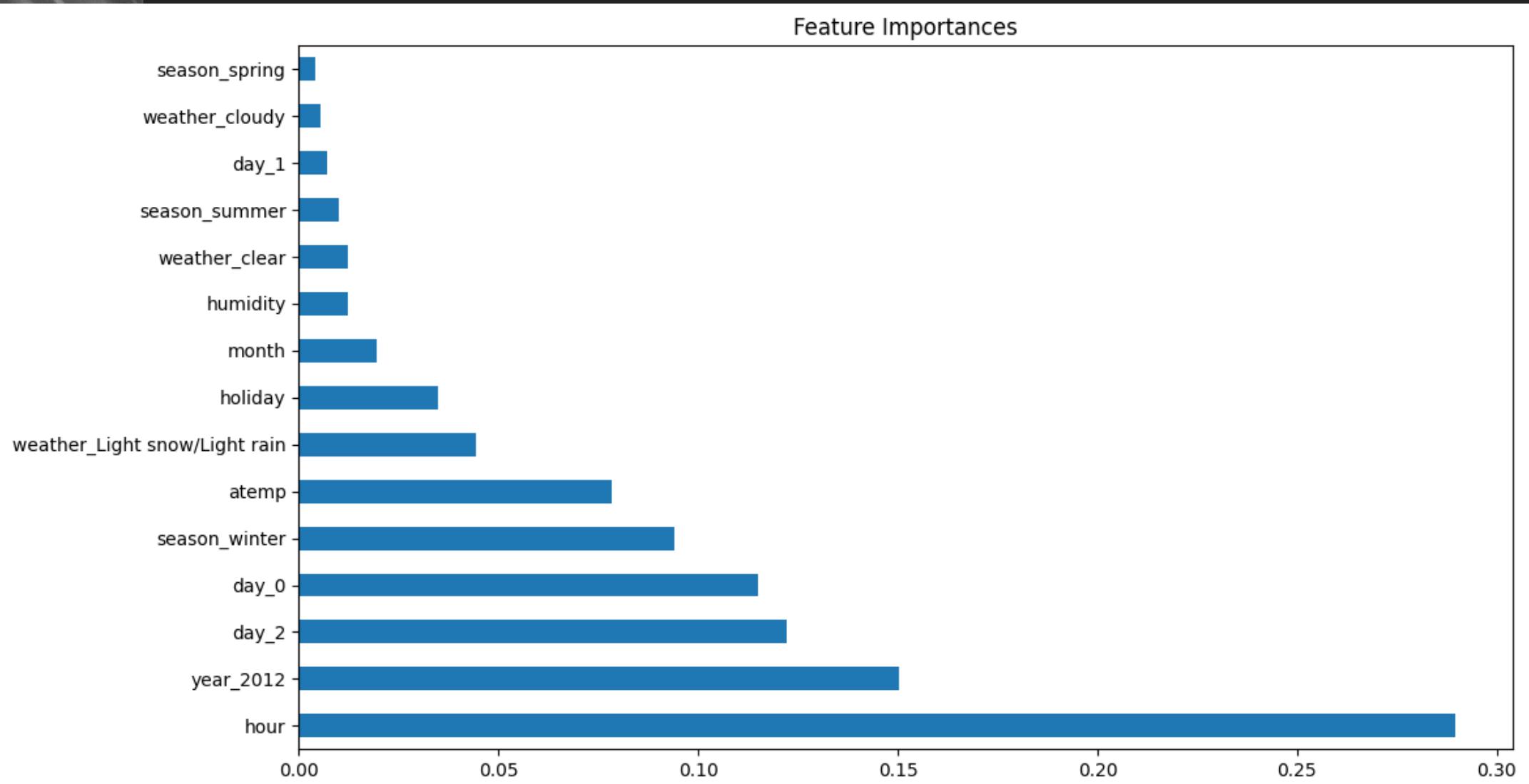


Count	Score MAE	Score MAPE
<=50	6.414451	0.479099
51-100	19.401112	0.272122
101-150	24.287318	0.195764
151-200	25.528128	0.147697
201-250	33.561909	0.150302
251-300	32.091666	0.116855
301-350	38.719513	0.119290
351-400	46.308589	0.123908
401-450	52.040149	0.123582
451-500	56.880694	0.119779
501-550	66.245597	0.127422
551-600	51.722427	0.090018
>600	63.180542	0.088679
All Count Range (Max 953)	26.174921	0.253480

BERDASARKAN HASIL PENGELOMPOKAN TARGET TERSEBUT TERLIHAT BAHWA TARGET DENGAN JUMLAH UNIT SEPEDA YANG DISEWA SAMPAI DENGAN 50 MEMILIKI NILAI MAPE YANG CUKUP BESAR (~50%), WALAUPUN PEMBAGIAN DATA TRAINING DAN TESTING UNTUK RANGE TERSEBUT SUDAH CUKUP BAIK YAITU TRAINING SET (2418) DAN TESTING SET (1008). DARI HASIL PENGELOMPOKKAN TERSEBUT, BISA DILIHAT LIMITASI MODEL DAPAT MEMPREDIKSI DENGAN BAIK UNTUK JUMLAH UNIT SEPEDA > 50 DENGAN HASIL NILAI EVALUASI METRIK YANG MASUK CUKUP BAIK.



FEATURE IMPORTANCES





CONCLUSION



Hyperparameter tuning terbaik dengan benchmark model XGBoost

- max_depth = 8
- learning_rate = 0.1
- n_estimators = 200

feature yang paling berpengaruh untuk count adalah:

- hour
- year
- season

Metrik evaluasi yang digunakan pada model ini adalah nilai MAE, MAPE, dan R2. Jika dilihat dari nilai MAPE yang dihasilkan oleh model setelah hyperparameter tuning (~25%), dapat disimpulkan bahwa bila model yang dibuat ini digunakan untuk memperkirakan jumlah unit sepeda yang harus disediakan pada rentang nilai yang sudah dilatih terhadap model (maksimal jumlah unit sepeda 970), maka perkiraan jumlahnya rata-rata akan meleset kurang lebih sebesar 25% dari jumlah yang seharusnya.

Namun dilihat dari hasil pengelompokan berdasarkan rentang target dan visualisasi residual, bisa dilihat limitasi model dapat memprediksi dengan baik untuk jumlah unit sepeda di atas 50 unit, dengan hasil nilai evaluasi metrik yang masih cukup baik dan reasonable. Tidak menutup kemungkinan bahwa prediksi yang dihasilkan bisa meleset lebih jauh. Bias tersebut dihasilkan karena terbatasnya feature pada dataset yang berkaitan dengan target (jumlah unit sepeda yang disewa) atau yang mampu merepresentasikan keadaan dimana calon pelanggan memutuskan untuk menggunakan jasa peminjaman sepeda seperti lokasi stasiun sepeda, jarak stasiun sepeda dengan perkantoran/sekolah/ruang publik, dll.

RECOMMENDATION

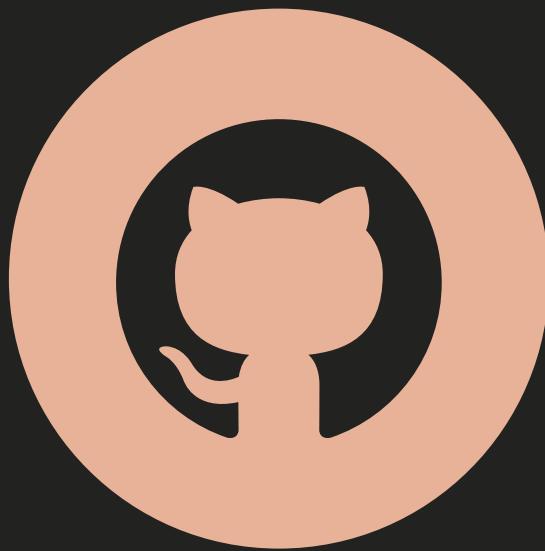


Adanya penambahan feature yang lebih berkorelasi terhadap target (`count`), seperti lokasi stasiun sepeda dan jarak stasiun sepeda dengan perkantoran/sekolah/ruang publik.

Adanya penambahan data, dataset yang digunakan hanya dalam rentang 1 tahun (2011-2012). Apabila ada penambahan rentang tahun data dalam dataset, hal itu tentu dapat membantu dalam meningkatkan prediksi dari model.

Model yang sudah dibuat ini dapat digunakan untuk mengembangkan pembuatan model yang lain seperti memprediksi total unit sepeda yang disewa pada lokasi tertentu. Hasil dari prediksi tersebut nantinya dapat dianalisa sebagai pertimbangan untuk menambah stasiun sepeda di lokasi-lokasi yang strategis.

THANK YOU



GITHUB

[https://github.com
/hrydwnt](https://github.com/hrydwnt)

Medium

MEDIUM

[https://medium.co
m/@hrydwnt22](https://medium.com/@hrydwnt22)