Machine Learning Module

Bike Sharing Prediction

PREPARED BY ADHANTO BAGASKORO

Outline

1. BUSINESS PROBLEM UNDERSTANDING
2. DATA UNDERSTANDING
3. DATA PREPROCESSING
4. MODELING
5. CONCLUSION
6. RECOMMENDATION

BUSINESS PROBLEM UNDERSTANDING

- Capital Bikeshare adalah sistem peminjaman sepeda di Amerika Serikat.
- Ini adalah generasi baru dari rental sepeda tradisional, di mana seluruh proses keanggotaan, penyewaan, dan pengembalian telah diotomasi.
- Pengguna dapat dengan mudah menyewa sepeda dari satu lokasi dan mengembalikannya di lokasi lain.
- Terdapat lebih dari 500 program bike-sharing di seluruh dunia, dengan lebih dari 500.000 sepeda.
- Capital Bikeshare sendiri memiliki lebih dari 600 stasiun dan sekitar 5.000 sepeda Capital Bikeshare yang tersebar di wilayah Washington.
- Sistem ini memiliki peran penting terhadap isu lalu lintas, lingkungan, dan kesehatan.

PROBLEM STATEMENT

- Jumlah sepeda yang tersedia harus cukup untuk memenuhi permintaan pengguna, tetapi tidak boleh terlalu banyak sehingga menyebabkan inefisiensi.
- Jika jumlah sepeda tidak cukup, pengguna mungkin tidak dapat menemukan sepeda yang tersedia saat mereka membutuhkannya. Ini dapat menyebabkan ketidakpuasan pelanggan dan hilangnya kepercayaan.
- Jika jumlah sepeda terlalu banyak, biaya manajemen, logistik, dan perawatan akan membengkak.

GOALS

- Menghasilkan prediksi jumlah sepeda yang tersedia yang akurat untuk setiap kondisi dan situasi.
- Menjaga efisiensi operasional biaya Capital Bikeshare.
- Memenuhi tuntutan kebutuhan pelanggan.

ANALISIS DATA

- Menemukan pola dari fitur-fitur yang ada, yang membedakan satu kondisi dengan yang lainnya.
- Mengidentifikasi bagaimana tiap fitur tersebut mempengaruhi jumlah unit sepeda yang perlu tersedia.

MODEL REGRESI

- Membangun model regresi untuk memprediksi jumlah unit sepeda yang perlu tersedia.
- Menggunakan model regresi untuk membantu dalam menentukan jumlah unit sepeda yang perlu disediakan oleh Capital Bikeshare.

METRIKS EVALUASI

- MAE: Rata-rata nilai absolut dari error. Semakin kecil nilai MAE, semakin akurat model dalam memprediksi jumlah unit sepeda.
- MAPE: Rata-rata persentase error yang dihasilkan oleh model regresi. Semakin kecil nilai MAPE, semakin akurat model dalam memprediksi jumlah unit sepeda.
- R-squared: Nilai yang menunjukkan seberapa baik model dapat merepresentasikan varians keseluruhan data. Semakin mendekati 1, semakin fit pula modelnya terhadap data observasi.

DATA UNDERSTANDING

INFORMASI DTASET – BIKE SHARING

Atribut	Tipe Data	Deskripsi
dteday	Objek	Tanggal
hum	Float	Kelembaban yang dinormalisasi (nilai dibagi 100)
weathersit	Integer	1: Cerah, Beberapa awan, Sebagian berawan, Sebagian berawan 2: Kabut + Berawan, Kabut + Awan pecah, Kabut + Beberapa awan, Kabut 3: Salju Ringan, Hujan Ringan + Badai petir + Awan berserak, Hujan Ringan + Awan berserak 4: Hujan Lebat + Butiran es + Badai petir + Kabut, Salju + Kabut
holiday	Integer	0: Bukan hari libur 1: Hari libur
season	Integer	1: Musim dingin 2: Musim semi 3: Musim panas 4: Musim gugur
atemp	Float	Suhu "terasa seperti" dalam Celsius
temp	Float	Suhu yang dinormalisasi dalam Celsius
hr	Integer	Jam (0 hingga 23)
casual	Integer	Jumlah pengguna sewa kasual
registered	Integer	Jumlah pengguna terdaftar
cnt	Integer	Jumlah total sepeda yang dipinjamkan, termasuk pengguna kasual dan terdaftar

DATA PREPROCESSING

1. PENYESUAIAN NAMA KOLOM DAN VALUE

	dteday	hum	weathersit	holiday	season	atemp	temp	hr	casual	registered	cnt
0	2011-12-09	0.62	1	0	4	0.3485	0.36	16	24	226	250
1	2012-06-17	0.64	1	0	2	0.5152	0.54	4	2	16	18
2	2011-06-15	0.53	1	0	2	0.6212	0.62	23	17	90	107
3	2012-03-31	0.87	2	0	2	0.3485	0.36	8	19	126	145
4	2012-07-31	0.55	1	0	3	0.6970	0.76	18	99	758	857

2.MENGUBAH TIPE DATA DAN MEMISAH DATA 'DATE”

```
RangeIndex: 12165 entries, 0 to 12164
Data columns (total 14 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   date        12165 non-null  datetime64[ns]
 1   humidity    12165 non-null  float64
 2   weather     12165 non-null  object
 3   holiday     12165 non-null  category
 4   season      12165 non-null  object
 5   atemp       12165 non-null  float64
 6   temp        12165 non-null  float64
 7   hour        12165 non-null  int64
 8   casual      12165 non-null  int64
 9   registered  12165 non-null  int64
 10  count       12165 non-null  int64
 11  year        12165 non-null  category
 12  month       12165 non-null  category
 13  day         12165 non-null  object
dtypes: category(3), datetime64[ns](1), float64(3), int64(4), object(3)
memory usage: 1.1+ MB
```

3.CHECKING MISSING VALUE DAN DATA DUPLIKAT

Pengecekan Nilai Hilang dan Data Duplikat

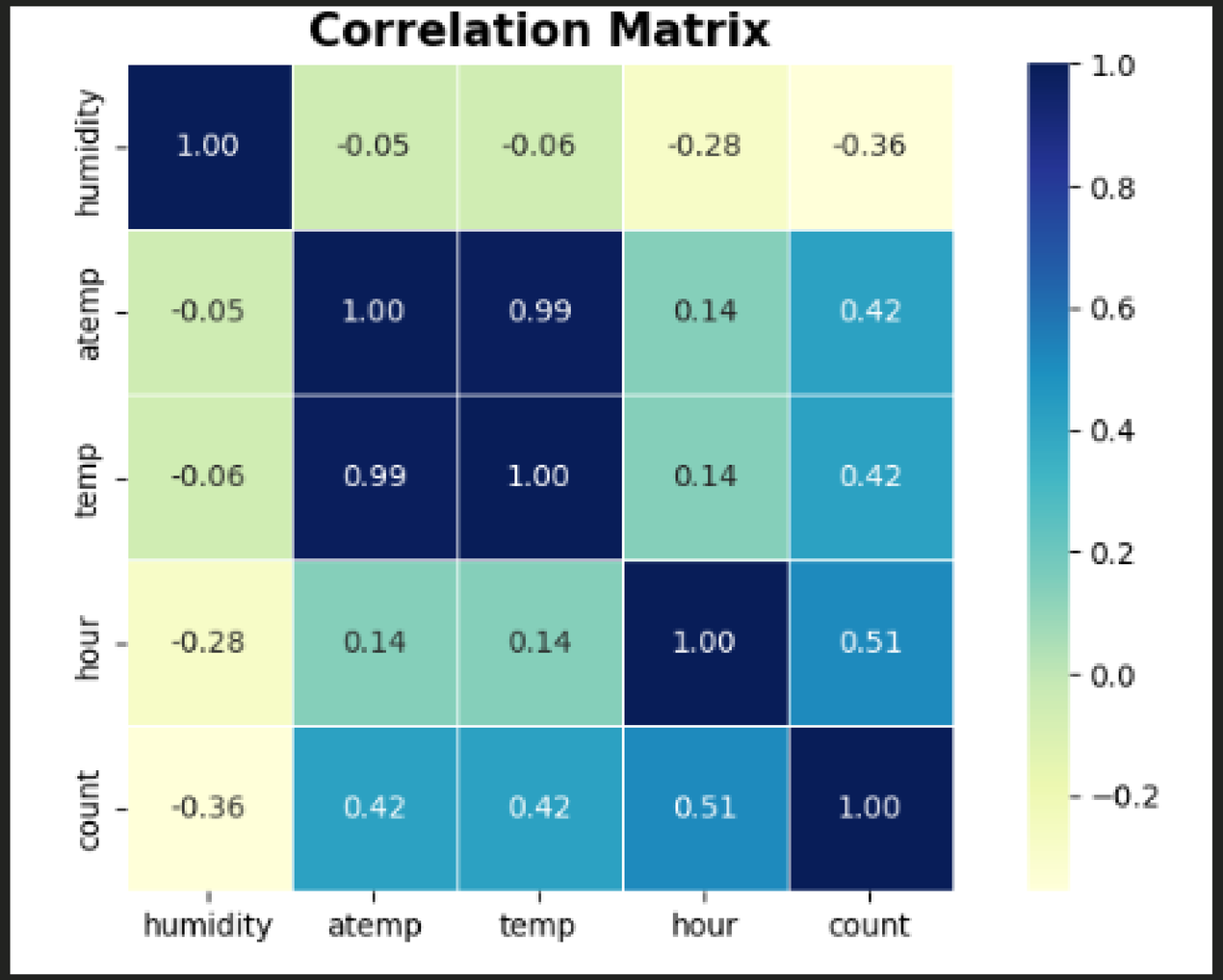
In [13]:
```
df_model.duplicated().any()
```

Out[13]: False

4.DROP COLUMN (FEATURE SELECTION)

DROP TEATURE DATE, CASUAL, DAN REGISTERED BERDASARKAN DOMAIN KNOWIEDGE.

5.CHECKING DATA CORRELATION (MATRIX DAN VIF SCORE)

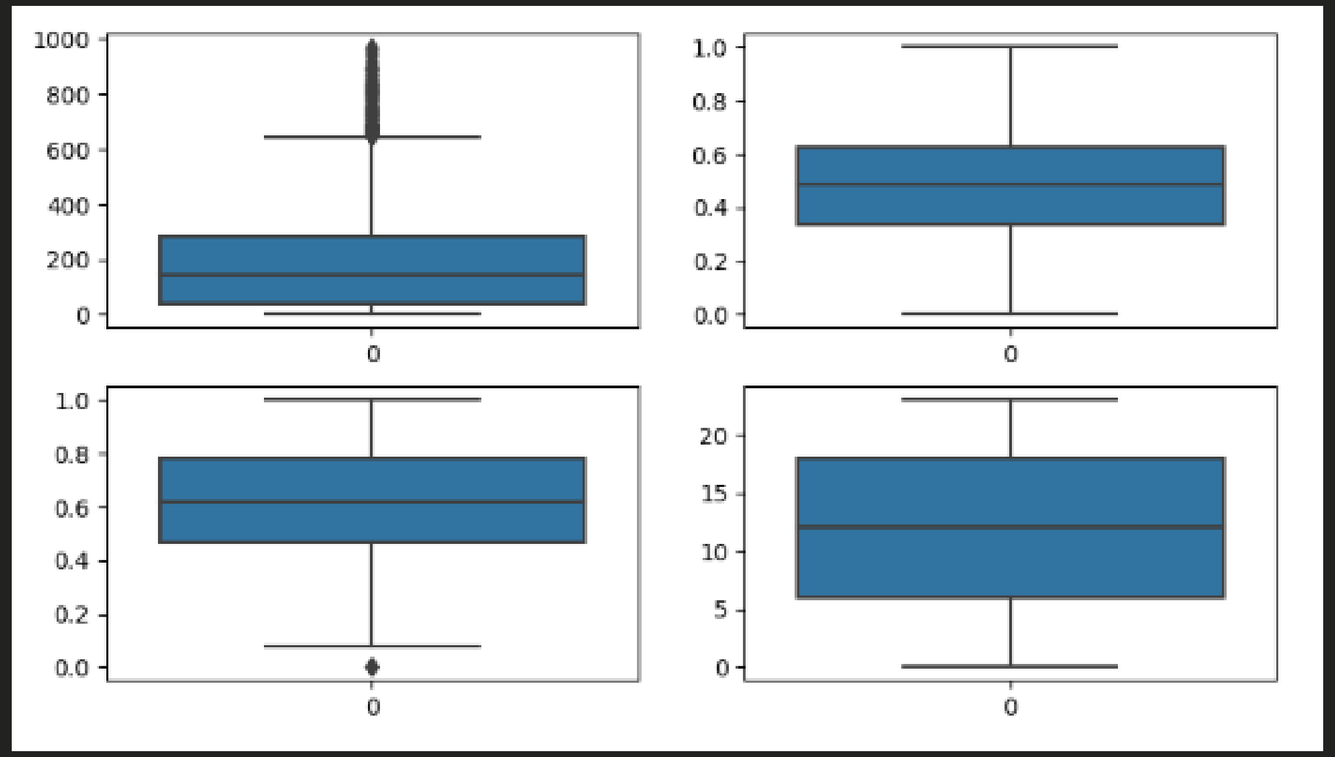


	variables	VIF
0	humidity	6.047913
1	atemp	338.317799
2	temp	306.108156
3	hour	3.849810
4	count	3.063218

	variables	VIF
0	humidity	5.571588
1	atemp	8.360153
2	hour	3.837270
3	count	3.063059

6.CHECKING OUTLIERS

- VARIABEL COUNT BERDISTRIBUSI MIRING KE KANAN.
- BATAS ATAS UNTUK VARIABEL COUNT ADALAH 645. TERDAPAT 338 BARIS DATA DENGAN NILAI LEBIH BESAR DARI 645.
- VARIABEL HUMIDITY MEMILIKI 14 BARIS DATA DENGAN NILAI 0.
- NILAI HUMIDITY TIDAK BOLEH 0.



7.CLEAN DATASET .

	humidity	weather	holiday	season	atemp	hour	count	year	month	day
0	0.81	clear	0	winter	0.2879	0	16	2011	1	Saturday
1	0.80	clear	0	winter	0.2727	1	40	2011	1	Saturday
2	0.80	clear	0	winter	0.2727	2	32	2011	1	Saturday
3	0.75	clear	0	winter	0.2879	3	13	2011	1	Saturday
4	0.75	clear	0	winter	0.2879	4	1	2011	1	Saturday

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 12151 entries, 0 to 12164
Data columns (total 10 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   humidity  12151 non-null  float64
 1   weather   12151 non-null  object
 2   holiday   12151 non-null  category
 3   season    12151 non-null  object
 4   atemp     12151 non-null  float64
 5   hour      12151 non-null  int64
 6   count     12151 non-null  int64
 7   year      12151 non-null  category
 8   month     12151 non-null  category
 9   day       12151 non-null  object
dtypes: category(3), float64(2), int64(2), object(3)
memory usage: 795.7+ KB
```

MODELLING

- ENCODING

ONE-HOT ENCODING DILAKUKAN PADA VARIABEL WEATHER, SEASON, DAN YEAR.
BINARY ENCODING DILAKUKAN PADA VARIABEL DAY.

- TRAIN AND TEST SPLITTING

DATA DIBAGI MENJADI 70% TRAIN SET DAN 30% TEST SET.

BASE MODEL: :
1.LINEAR REGRESSION
2.K-NEAREST NEIGHBORS REGRESSOR
3. DECISION TREE REGRESSOR

ENSEMBIE METHOD MODEL:
1.RANDOM FOREST REGRESSOR
2. GRADIENT BOOSTING REGRESSOR
3. EXTREME GRADIENT BOOSTING REGRESSOR (XGBOOST)

CHOOSING BENCHMARK MODEL (USING DATA TRAIN)

	Model	MAE	MAPE	R-squared
0	Linear Regression	-106.333268	-1.389270	0.188748
1	KNN Regressor	-46.134930	-0.454068	0.824984
2	DecisionTree Regressor	-45.672462	-0.512351	0.797346
3	RandomForest Regressor	-35.522202	-0.349619	0.889855
4	Gradient Boosting	-55.468159	-0.487001	0.756239
5	XGBoost Regressor	-29.914250	-0.279923	0.923665

Hasil evaluasi model

- Model XGBoost memiliki skor terbaik berdasarkan evaluasi menggunakan K-Fold Cross Validation.
- Skor model XGBoost:
  - MAE (Mean Absolute Error): 29.91
  - MAPE (Mean Absolute Percentage Error): -27.26%
  - R-squared: 0.92

DARI HASIL EVALUASI MENGGUNAKAN K-FOLD CROSS VALIDATION, XGBOOST MEMILIKI SKOR TERBAIK DENGAN NILAI MAE 29.91, MAPE ~27X, DAN R-SQUARED 0.92.

EXTREME GRADIENT BOOSTING

XGBoost

- XGBoost adalah model ensemble berbasis pohon keputusan yang menggunakan teknik boosting untuk meningkatkan akurasi.
- XGBoost menggunakan bobot untuk mengontrol pentingnya setiap variabel independen dalam prediksi.
- XGBoost merupakan model non-interpretable, artinya kita tidak dapat mengetahui variabel mana yang salah diprediksi.
- XGBoost memiliki kinerja yang baik pada berbagai jenis data.

- XGBOOST ADALAH MODEL ENSEMBLE YANG MENGGABUNGKAN SEJUMLAH DECISION TREE.
- DECISION TREE DITAMBAHKAN SECARA SEKUENSIAL, DENGAN BOBOT YANG DISESUAIKAN UNTUK MENINGKATKAN AKURASI PREDIKSI.
- XGBOOST ADALAH MODEL NON-INTERPRETABLE, TETAPI MEMILIKI KINERJA YANG BAIK UNTUK BERBAGAI JENIS DATA.

EXTREME GRADIENT BOOSTING

	MAE	MAPE	R-squared
XGB	28.692625	0.262615	0.937314

- XGBOOST MEMILIKI PERFORMA YANG BAIK PADA TEST SET, DENGAN NILAI MAE DAN MAPE YANG MENURUN, SERTA NILAI R-SQUARED YANG MENINGKAT.
- NILAI MAE DAN MAPE MENURUN DARI 29.91 DAN -27.26% MENJADI 28.69 DAN -26.5X%.
- NILAI R-SQUARED MENINGKAT DARI 0.92 MENJADI 0.93.

HYPERPARAMETER TUNING (GRID SEARCH)

Parameter	Value
max_depth	2,3,4,5,6,7,8,9
learning_rate	0.1, 0.001, 0.0001, 0.2, 0.3, 0.5, 0,7
n_estimators	200, 220, 240, 260, 280, 300

- Tuning hyperparameters adalah proses penyesuaian parameter model untuk meningkatkan kinerjanya.
- Pada XGBoost, terdapat sejumlah hyperparameters yang dapat dituning, seperti:
  - `max_depth` : kedalaman pohon
  - `learning_rate` : ukuran step pada setiap iterasi
  - `n_estimators` : jumlah pohon
- Tuning hyperparameters dapat membantu mencegah overfitting, meningkatkan akurasi, dan menjaga efisiensi.

HYPERPARAMETER TUNING RESULT

Parameter	Best Value
nax_depth	8
learning_rate	0.1
n_estimators	260

PREDICT TO TEST SET USING BEST PARAMETERS

	MAE	MAPE	R-squared
XGB	26.238034	0.25205	0.945908

PERFORMANCE COMPARISON

BERIKUT ODALAH KOMPARASI PERTORMA MODEL XOBOOST SEBELUM DAN SESUDAH DILAKUKAN HYPERPARAMETER TUNING:

```
score_before_tuning
```

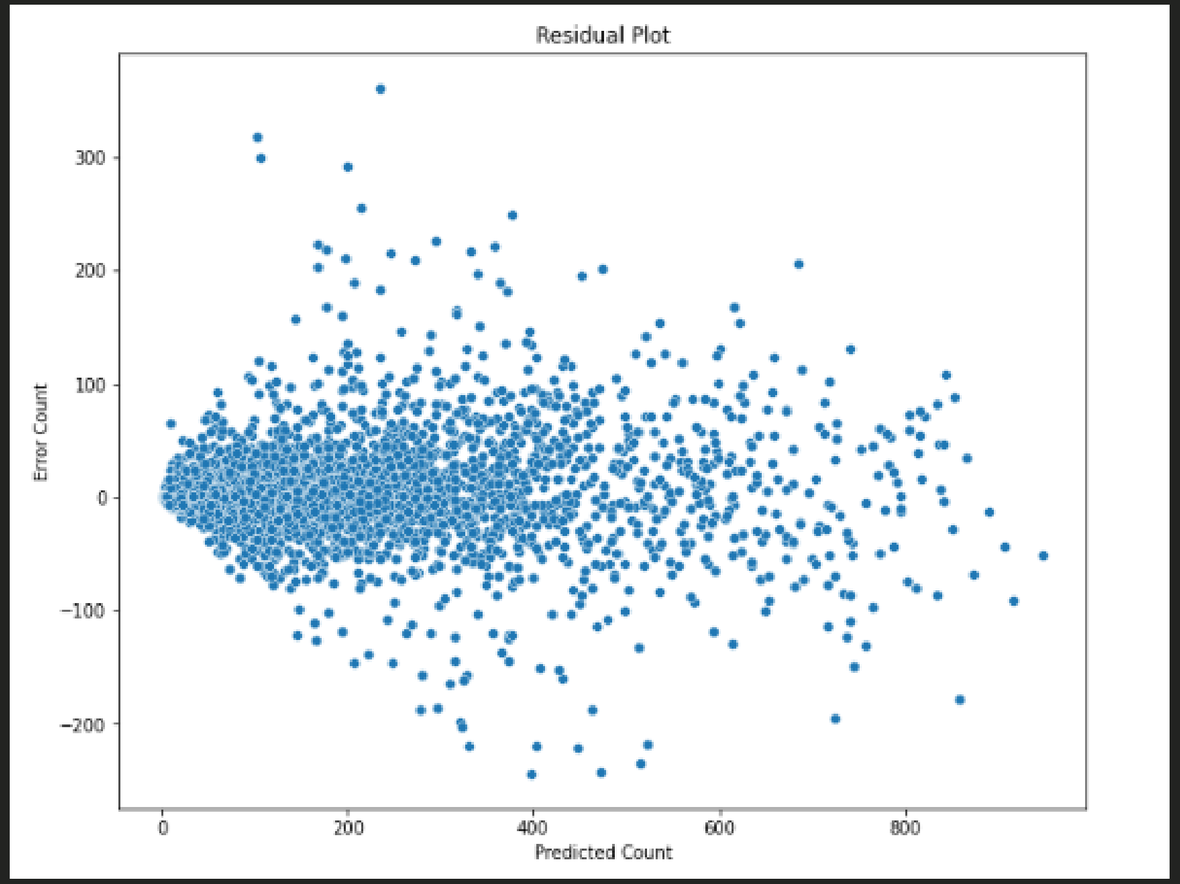
	MAE	MAPE	R-squared
XGB	28.692625	0.262615	0.937314

```
score_after_tuning
```

	MAE	MAPE	R-squared
XGB	26.238034	0.25205	0.945908

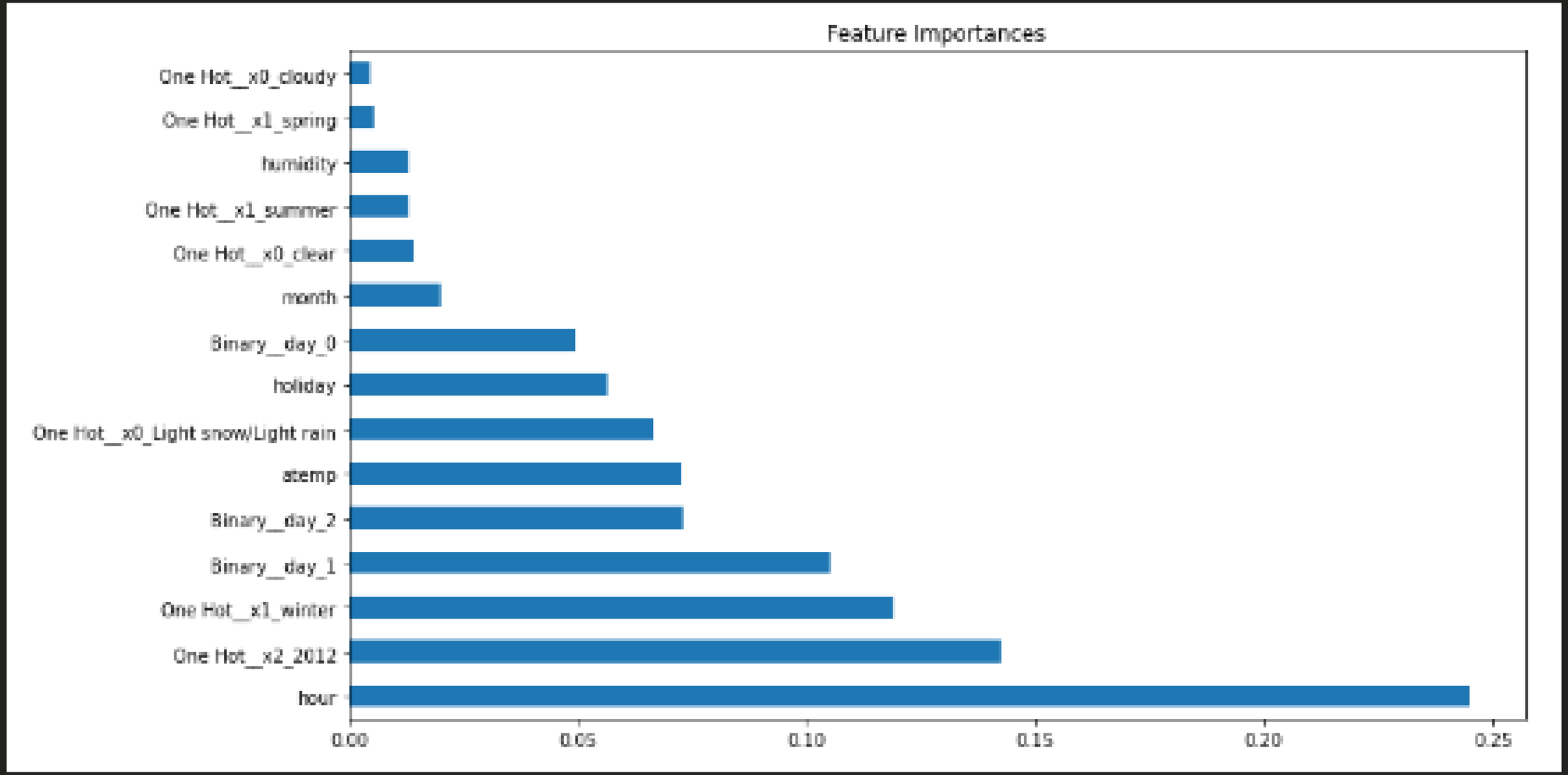
MODEL LIMITATION

	Score MAE	Score MAPE
<=50	7.131987	0.502871
51-100	18.453602	0.255349
101-150	23.727912	0.193825
151-200	26.044959	0.151437
201-250	29.457202	0.131500
251-300	30.723859	0.111833
301-350	39.700542	0.123305
351-400	43.271628	0.115724
401-450	49.671722	0.118235
451-500	62.212817	0.131262
501-550	63.140174	0.120219
551-600	54.628588	0.095113
>600	60.368069	0.085163
All Count Range (Max 953)	26.238034	0.252050



- UNTUK TARGET DENGAN JUMLAH UNIT SEPEDA YANG DISEWA SAMPAI DENGAN 50, NILAI MAPE CUKUP BESAR (~50%).
- PEMBAGIAN DATA TRAINING DAN TESTING UNTUK RANGE TERSEBUT SUDAH CUKUP BAIK.
- MODEL DAPAT MEMPREDIKSI DENGAN BAIK UNTUK JUMLAH UNIT SEPEDA DI ATAS 50 UNIT.

FEATURE IMPORTANCES



CONCLUSION & REKOMENDATION

CONCLUSION

- Hasil hyperparameter tuning untuk XGBoost adalah:
  - `max_depth` : 8
  - `learning_rate` : 0.1
  - `n_estimators` : 260
- Feature yang paling berpengaruh terhadap jumlah unit sepeda yang disewa adalah:
  - `hour`
  - `year`
  - `season`
- Nilai MAPE yang dihasilkan oleh model adalah (~25%).
- Model dapat memprediksi dengan baik untuk jumlah unit sepeda di atas 50 unit.
- Model dapat menghasilkan prediksi yang meleset lebih jauh untuk jumlah unit sepeda di bawah 50 unit.

- Nilai MAPE yang sebesar ~25% menunjukkan bahwa model memiliki kesalahan prediksi yang cukup besar.
- Feature hour, year, dan season menunjukkan bahwa waktu, tahun, dan musim memiliki pengaruh yang besar terhadap jumlah unit sepeda yang disewa.
- Model yang dapat memprediksi dengan baik untuk jumlah unit sepeda di atas 50 unit menunjukkan bahwa model tersebut dapat mempelajari pola data dengan baik untuk jumlah unit sepeda yang lebih besar.
- Prediksi model yang dapat meleset lebih jauh untuk jumlah unit sepeda di bawah 50 unit menunjukkan bahwa model tersebut memiliki bias terhadap jumlah unit sepeda yang lebih kecil. 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.

RECOMMENDATION

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

2. 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.

3. Model yang sudah dibuat ini dapat digunakan untuk mengembangkan pembuatan model yang lain. Seperti memprediksi total unit sepeda yang disewa pada lokasi tertentu. Dimana nantinya dapat dianalisa sebagai pertimbangan untuk menambah stasiun sepeda di lokasi-lokasi yang strategis.

Thank You!