UAS Mathematical Tools For Data Science

Resume Model Regression

Disusun oleh :

Muhamad Irvan Dandung

(41519110136)

Dosen Pengampu :

Ida Nurhaida, Dr,MT.

Universitas Mercu Buana – Menteng

Jl. Menteng Raya No.29, RT.1/RW.10, Kb. Sirih, Kec. Menteng, Kota Jakarta Pusat, Daerah Khusus Ibukota Jakarta 10340.

BAB I
PENDAHULUAN

Prediksi pada dasarnya merupakan dugaan atau prediksi mengenai terjadinya suatu kejadian atau peristiwa di waktu yang akan datang. Prediksi bisa bersifat kualitatif (tidak berbentuk angka) maupun kuantitatif (berbentuk angka). Prediksi kualitatif sulit dilakukan untuk memperoleh hasil yang baik karena variabelnya sangat relatif sifatnya. Prediksi kuantitatif dibagi dua yaitu: prediksi tunggal (point prediction) dan prediksi selang (interval prediction). Prediksi tunggal terdiri dari satu nilai, sedangkan prediksi selang terdiri dari beberapa nilai, berupa suatu selang (interval) yang dibatasi oleh nilai batas bawah (prediksi batas bawah) dan batas atas (prediksi tinggi) . Prediksi berfungsi untuk membuat suatu rencana kebutuhan (demand) yang harus dibuat yang dinyatakan dalam kuantitas (jumlah) sebagai fungsi dari waktu. Prediksi dilakukan dalam jangka panjang (long term). Prediksi yang berkaitan dengan pernyataan (1) what will be demanded, (2) how many, dan (3) when it should be supplied? Prediksi sangat diperlukan dengan melakukan perbandingan antara kebutuhan yang diramalkan dengan yang sebenarnya.

Regresi Linear Sederhana atau sering disingkat dengan SLR (Simple Linear Regression) juga merupakan salah satu metode statistik yang dipergunakan dalam produksi untuk melakukan peramalan ataupun prediksi tentang karakteristik kualitas maupun kuantitas.

1. Tujuan Penelitian

Adapun tujuan adanya penelitian atau pengerjaan tugas besar 2 ini yaitu:

1. Mempelajari & menerapkan Metode Simple Linear Regression(SLR).
2. Menganalisa pengaruh fitur-fitur rumah terhadap harga rumah tersebut.
3. Memprediksikan harga beberapa rumah lainnya.

2. Metode Penelitian

Data mining adalah penambangan atau penemuan informasi baru dengan mencari pola atau aturan tertentu dari sejumlah data yang sangat besar. Data mining juga disebut sebagai serangkaian proses untuk menggali nilai tambah berupa pengetahuan yang selama ini tidak diketahui secara manual dari suatu kumpulan data. Data mining, sering juga disebut sebagai Knowledge Discovery In Database (KDD). KDD adalah kegiatan yang meliputi pengumpulan, pemakaian data, histori untuk menemukan keteraturan, pola atau hubungan dalam set data berukuran besar. Data mining adalah kegiatan menemukan pola yang menarik dari data dalam jumlah besar, data dapat disimpan dalam database, data warehouse, atau penyimpanan informasi lainnya. Data mining berkaitan dengan bidang ilmu-ilmu lain, seperti database system, data warehouse, statistical, machine learning, information retrieval, dan komputasi tingkat tinggi. Selain itu, data mining didukung oleh ilmu lain seperti neural network, pengenalan pola, spatial data analysis, image database, signal processing.

A. Machine Learning (ML)

Teknologi machine learning (ML) adalah mesin yang dikembangkan untuk bisa belajar dengan sendirinya tanpa arahan dari penggunanya. Pembelajaran mesin

dikembangkan berdasarkan disiplin ilmu lainnya seperti statistika, matematika dan data mining sehingga mesin dapat belajar dengan menganalisa data tanpa perlu di program ulang atau diperintah. Dalam hal ini machine learning memiliki kemampuan untuk memperoleh data yang ada dengan perintah ia sendiri. ML juga dapat mempelajari data yang ada dan data yang ia peroleh sehingga bisa melakukan tugas tertentu. Tugas yang dapat dilakukan oleh ML pun sangat beragam, tergantung dari apa yang ia pelajari.
Ada 3 jenis Machine Learning (ML) yang sekarang lagi populer, yaitu :

1. Supervised Learning,
2. Unsupervised Learning,
3. dan Reinforcement Learning.

B. Simple Linear Regression(SLR)

Simple linear Regression hanya mempunyai 1 independent variabel (x). Walaupun sederhana, algoritma ini merupakan salah satu algoritma yang sangat populer karena simple tapi powerful. Untuk Menganalisa pengaruh fitur-fitur rumah terhadap harga rumah dan memprediksikan harga beberapa rumah lainnya. Menggunakan algoritma prediksi regression atau Regresi linear sederhana, Regresi linear sederhana adalah analisis regresi yang melibatkan hubungan antara satu variabel tak bebas dihubungkan dengan satu variabel bebas. Regresi linier juga merupakan metode statistik yang berfungsi untuk menguji sejauh mana hubungan sebab-akibat antara variabel faktor penyebab (x) terhadap variabel akibatnya. Faktor penyebab pada umumnya dilambangkan dengan X sedangkan variabel akibat dilambangkan dengan Y. Regresi linear sederhana atau sering disingkat dengan SLR (Simple Linear Regression) juga merupakan salah satu metode statistik yang dipergunakan dalam produksi untuk melakukan peramalan atau pun prediksi tentang karakteristik kualitas maupun kuantitas. Persamaan umum metode regresi linier sederhana dalam penelitian ini adalah: $Y = a + b(X)$ Keterangan: a = Konstanta b = Koefisien regresi Y= Variabel dependen (variabel tak bebas) X = Variabel independen (variabel bebas). Menentukan koefisien persamaan a dan b dapat dengan menggunakan metode kuadrat terkecil, yaitu cara yang dipakai untuk menentukan koefisien persamaan dan dari jumlah pangkat dua (kuadrat) antara titik-titik dengan garis regresi yang dicari yang terkecil.

BAB II
ISI

1. Konfigurasi Parameter

A. Menyediakan Data Set

Untuk data set sendiri itu diambil dari kc_house_data.csv. Untuk mengambil data tersebut ada beberapa hal yang dilakukan yaitu mendownload file kc_house_data.csv yang telah disediakan sehingga dapat di inject ke dalam project.

```
[1] !wget -O /content/sample_data/kc_house_data.csv https://raw.githubusercontent.com/irvandandung/kc_house_data_regression/master/kc_house_data.csv

--2021-08-24 13:47:13--  https://raw.githubusercontent.com/irvandandung/kc_house_data_regression/master/kc_house_data.csv
Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 185.199.108.133, 185.199.109.133, 185.199.110.133, ...
Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|185.199.108.133|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 2515206 (2.4M) [text/plain]
Saving to: '/content/sample_data/kc_house_data.csv'

/content/sample_dat 100%[===================>]   2.40M  --.-KB/s    in 0.07s

2021-08-24 13:47:14 (35.8 MB/s) - '/content/sample_data/kc_house_data.csv' saved [2515206/2515206]
```

B. Package yang digunakan

- Numpy => digunakan untuk mengubah data pada dataframe menjadi data array atau data list.
- Matplotlib.pyplot => digunakan untuk menampilkan data pada dataframe menjadi bentuk visual sesuai kebutuhan.
- Pandas => digunakan untuk mengolah data csv menjadi dataframe.
- Sklearn => digunakan memproses data list menjadi normal / digunakan untuk menormalkan data.
- Seaborn => kegunaannya sama seperti Matplotlib.pyplot, package ini digunakan untuk menampilkan data dalam bentuk visual.

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
```

C. Pembersihan Data

Untuk pembersihan data sendiri hal yang pertama dilakukan adalah menentukan dari sekian banyak column, column mana saja yang dijadikan atribute, lalu setelah itu

yang dilakukan selanjutnya yaitu cek apakah ada data yang kosong, lalu terakhir isi data yang kosong dan lakukan pembuangan column yang tidak dipakai.

- Cek ada data yang null atau kosong tidak.

```
dataframe.isnull().sum()
```

```
id               0
date             0
price            0
bedrooms         0
bathrooms        0
sqft_living      0
sqft_lot         0
floors           0
waterfront       0
view             0
condition        0
grade            0
sqft_above       0
sqft_basement    0
yr_built         0
yr_renovated     0
zipcode          0
lat              0
long             0
sqft_living15    0
sqft_lot15       0
dtype: int64
```

- Karena tidak ada data yang kosong maka, masuk ke tahap pembersihan column yang tidak dipakai.

```
dataframe.drop(['date','id', 'zipcode','view','lat','long','sqft_living15', 'sqft_lot15'], axis=1, inplace=True)
dataframe.head(2)
```

	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	condition	grade	sqft_above	sqft_basement	yr_built	yr_renovated
0	221900.0	3	1.00	1180	5650	1.0	0	3	7	1180	0	1955	0
1	538000.0	3	2.25	2570	7242	2.0	0	3	7	2170	400	1951	1991

Dari data di atas ada beberapa column data yang dibuang yaitu date, id, zipcode, view, lat, long, sqft_livings, sqft_lotis. Alasan column data di atas dibuang dikarenakan ada beberapa column yang tidak mempengaruhi dari harga atau price sebuah rumah dan ada beberapa column data yang datanya tidak diketahui kegunaannya buat apa.

D. **Feature Set**

Setelah data sudah siap, maka langkah selanjutnya adalah menentukan x dan y dimana x sebagai data yang akan di train dan y adalah label. Jadi untuk x merupakan data normal dari data column bedrooms, bathrooms, sqft_living, sqft_lot, floors, waterfront, condition, grade, sqft_above, sqft_basement, yr_built, yr_renovated, sedangkan untuk y sebagai label di ambil dari data priced karena data yang dianalisa yaitu kemungkinan bertahan hidup penumpang yang ada di tata tersebut. Baru ketika x dan y sudah didapatkan sebagai argumen maka barulah dapat dilakukan model regresi.

```
x = dataframe.drop(columns='price')
y = dataframe['price']
print(x)
print(y)
```

```
       bedrooms  bathrooms  sqft_living  ...  sqft_basement  yr_built  yr_renovated
0             3       1.00         1180  ...              0      1955             0
1             3       2.25         2570  ...            400      1951          1991
2             2       1.00          770  ...              0      1933             0
3             4       3.00         1960  ...            910      1965             0
4             3       2.00         1680  ...              0      1987             0
...         ...        ...          ...  ...            ...       ...           ...
21608         3       2.50         1530  ...              0      2009             0
21609         4       2.50         2310  ...              0      2014             0
21610         2       0.75         1020  ...              0      2009             0
21611         3       2.50         1600  ...              0      2004             0
21612         2       0.75         1020  ...              0      2008             0

[21613 rows x 12 columns]
0        221900.0
1        538000.0
2        180000.0
3        604000.0
4        510000.0
           ...
21608    360000.0
21609    400000.0
21610    402101.0
21611    400000.0
21612    325000.0
Name: price, Length: 21613, dtype: float64
```

2. Penentuan Nilai Skor Modelling

Setelah melakukan konfifurasi parameter langkah selanjutnya yaitu mencari skor akurasi dengan menggunakan algoritma simple linear regression.

1. Lakukan split data kita menjadi training and testing dengan porsi 80:20.
2. Lalu buat object linear regresi.

```
✓ [18] from sklearn.linear_model import LinearRegression
0s
```

```
✓ [19] lin_reg = LinearRegression()
0s
```

3. Training the model menggunakan training data yang sudah displit sebelumnya.

```
[20] lin_reg.fit(x_train, y_train)

     LinearRegression(copy_X=True, fit_intercept=True, n_jobs=None, normalize=False)
```

```
[21] print(lin_reg.coef_)
     print(lin_reg.intercept_)

     [[-4.20562884e+04  4.45981163e+04  1.25237482e+02 -2.63652763e-01
        2.33522544e+04  6.90391774e+05  1.93129435e+04  1.26801801e+05
        5.58904140e+01  6.93470678e+01 -3.67677858e+03  1.12043786e+01]]
     [6380227.61732255]
```

4. Cari tahu accuracy score dari model kita menggunakan testing data yang sudah di split sebelumnya.

```
[ ] y_pred = lin_reg.predict(x_test)
```

```
[ ] lin_reg.score(x_test, y_test)

    0.645910558180731
```

5. Model mendapatkan accuracy score sebesar 66.64% dari hasil di atas.

3. Prediksi

```
priceItemOrigin = 538000.0
predict = lin_reg.predict([[3,2,2570,7242,2.0,0,3,7,2170,400,1951,1991]])
print('harga predict:', predict[0])
print('selisih price asli dengan price predict: ', predict[0] - priceItemOrigin)
```

```
harga predict: 667638.8652519565
selisih price asli dengan price predict:  129638.86525195651
```

Dari hasil prediksi di atas, digunakan sebuah sampel data yang ada di atas dimana data rumah tersebut memiliki harga asli 538.000. Lalu dari hasil prediksi didapatkan harga prediksinya yaitu sekitar 667.683. Sehingga selisih price asli dengan price prediksi didapatkan yaitu sekitar +100.000.

BAB III
KESIMPULAN

Dari pembahasan yang sudah dijelaskan di bab 3 dapat disimpulkan bahwa salah satu algoritma yang dapat digunakan untuk menganalisa suatu harga / nilai dari rumah dengan data fitur-fitur rumah yang ada tersebut adalah dengan menggunakan Algoritma regresi. Walaupun disini algoritma yang digunakan yaitu algoritma regresi yang paling simple, skor yang dihasilkan mencapai 64 % dan hasil test yang didapatkan yaitu harga prediksi sekitar 100.000 an lebih mahal dibanding harga asli yang ada didalam data csv.

REFERENSI

- Larose, Daniel T., 2005, "Discovering Knowledge in Data: An Introduction to Data Mining", John Willey & Sons, Inc, New Jersey.
- McLeod, Raymond, 1995, "Sistem Informasi Manajemen",PT. Tema Baru, Klaten
- Santosa, Budi, 2007, "Data Mining Teknik Pemanfaatan Data untuk Keperluan Bisnis", Graha Ilmu,Yogyakarta.
- Susanto, Sani, dan Suryadi, Dedy, 2010, "Pengantar Data Mining" , Penerbit Andi Yogyakarta, Yogyakarta.