Laporan UTS - IBDA3111

Calvin Institute of Technology

Semester ganjil 2022/2023

Oleh

James Patrick Oentoro / 202000241 / IT & Big Data Analytics

Rental Apartemen Amsterdam

Latar belakang : Amsterdam adalah kota terbesar di Belanda, sebagai perantau ataupun orang awam yang sedang mencari tempat tinggal dengan menyewa apartemen, susah bagi kita untuk menentukan apakah harga yang diberikan oleh sebuah apartemen sudah sesuai. Maka dari itu, melalui data-data yang sudah dikumpulkan, kita bisa memprediksi harga sebuah apartemen dengan data title, price, area, volume, type, year, no_rooms, no_bed, no_floor, balcony, garden, dan interior. Dengan memanfaatkan model machine learning dan mengoptimalkan model melalui pembersihan data, imputasi data, dan seleksi fitur diperlukan agar data lebih akurat.

Data diambil melalui hasil crawling web (pararius.com) menggunakan library python yaitu selenium, data-data yang diambil meliputi 2080 barius dengan 12 fitur, yaitu:

1. Title : Nama apartemen yang disewakan.
2. Price : Harga penyewaan apartemen tersebut per bulan.
3. Area : Luas apartemen dalam satuan meter kubik.
4. Type : Tipe bangunan apakah itu apartemen atau rumah.
5. Year : Tahun dibangunnya apartemen tersebut.
6. No_rooms : Jumlah ruangan yang ada pada apartemen tersebut.
7. No_bed : Jumlah kamar tidur yang ada pada apartemen.
8. No_bath : Jumlah kamar mandi yang ada pada apartemen.
9. No_floor : Jumlah lantai yang ada pada apartemen.
10. Balcony : Apakah terdapat balkon pada apartemen.
11. Garden : Apakah terdapat taman dan luasnya.
12. Interior : Apakah interior berperabot, berlapis kain, atau tidak.

Beberapa teknik prapemrosesan dan rekayasa data saya lakukan untuk menyesuaikan data agar model lebih baik, pertama menyesuaikan data sehingga informasi sesuai dengan yang dibutuhkan mis. Mengubah kolom area yang sebelumnya object menjadi numeric dalam satuan meter kubik. Selanjutnya saya melakukan pembersihan data dengan menghapus data terduplikasi, *single value*, dan menghapus kolom yang memiliki jumlah nilai unik sama dengan data. Kemudian saya menganalisa distribusi data agar saya mendapat gambaran teknik-teknik apa yang sesuai dengan data kedepannya. Saya juga melakukan encoding menggunakan BinaryEncoder untuk mengubah data-data categorical menjadi numeric agar dapat melatih model dan

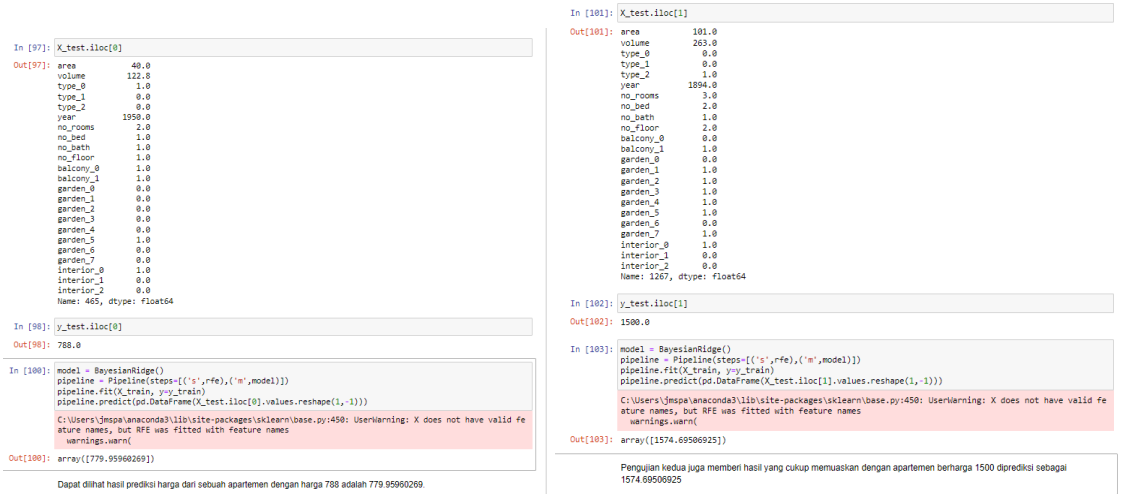
melakukan inputasi data dengan metode KNN imputer karena data yang kosong dapat diperkirakan berdasarkan hubungannya dengan data-data yang lain (cont. bila rumah memiliki luas yang besar maka cenderung rumah tersebut dibangun pada tahun tertentu).

Kemudian saya menghapus data-data outlier dengan metode Automatic Outlier Detection, karena data tidak terdistribusi secara Gaussian dan tidak terdapat data yang terpencil terlalu jauh. Tidak hanya itu, saya juga membuat model feature importance menggunakan random forest untuk regresi untuk melihat seberapa berpengaruhnya dari masing-masing fitur tergadap model. Kemudian saya melakukan Feature Selection dengan metode RFE karena lebih fleksibel dan dapat menggunakan algoritma-algoritma yang berbeda untuk mengindikasikan variable importance. Hingga akhirnya saya menggunakan beberapa model machine learning untuk menguji model mana yang paling sesuai dengan data.

Solusi yang saya dapat berbentuk model machine learning yang sudah diukur mean absolute error-nya (MAE). Beberapa model saya coba untuk menemukan model terbaik, yaitu, LinearRegression, RANSACRegressor, RandomForestRegressor, KernelRidge, BayesianRidge, SGDRegressor, ElasticNet, dan GradientBoostingRegressor. Seluruh model kecuali SGDRegressor, menunjukkan nilai MAE yang mirip yaitu antara -432.635 dengan standar deviasi diantara 28.479. Model ini bertujuan untuk memprediksi harga sebuah apartemen berdasarkan data-data yang sudah disebutkan sebelumnya, untuk membantu pengguna dalam pertimbangan.

Melalui model ini, kita sebagai orang awam maupun peratau yang ingin melakukan transaksi penyewaan apartemen dapat mempertimbangkan harga yang ditawarkan penjual dengan harga yang ditentukan oleh model. Meski model machine learning ini belum sepenuhnya akurat, namun dapat dijadikan pertimbangan dan dikembangkan lagi.

Contoh prediksi :



```
In [97]: X_test.iloc[0]
Out[97]: area           40.0
         volume        122.8
         type_0          1.0
         type_1          0.0
         type_2          0.0
         year         1950.0
         no_rooms        2.0
         no_bed          1.0
         no_bath         1.0
         no_floor        1.0
         balcony_0       1.0
         balcony_1       1.0
         garden_0        0.0
         garden_1        0.0
         garden_2        0.0
         garden_3        0.0
         garden_4        0.0
         garden_5        1.0
         garden_6        0.0
         garden_7        0.0
         interior_0      1.0
         interior_1      0.0
         interior_2      0.0
         Name: 465, dtype: float64

In [98]: y_test.iloc[0]
Out[98]: 788.0

In [100]: model = BayesianRidge()
          pipeline = Pipeline(steps=[('s',rfe),('m',model)])
          pipeline.fit(X_train, y=y_train)
          pipeline.predict(pd.DataFrame(X_test.iloc[0].values.reshape(1,-1)))
          C:\Users\jmspa\anaconda3\lib\site-packages\sklearn\base.py:450: UserWarning: X does not have valid fe
          ature names, but RFE was fitted with feature names
            warnings.warn(
Out[100]: array([779.95960269])
```

Dapat dilihat hasil prediksi harga dari sebuah apartemen dengan harga 788 adalah 779.95960269.

```
In [101]: X_test.iloc[1]
Out[101]: area           101.0
          volume         263.0
          type_0           0.0
          type_1           0.0
          type_2           1.0
          year          1894.0
          no_rooms         3.0
          no_bed           2.0
          no_bath          1.0
          no_floor         2.0
          balcony_0        0.0
          balcony_1        1.0
          garden_0         0.0
          garden_1         1.0
          garden_2         1.0
          garden_3         1.0
          garden_4         1.0
          garden_5         1.0
          garden_6         0.0
          garden_7         1.0
          interior_0       1.0
          interior_1       0.0
          interior_2       0.0
          Name: 1267, dtype: float64

In [102]: y_test.iloc[1]
Out[102]: 1500.0

In [103]: model = BayesianRidge()
          pipeline = Pipeline(steps=[('s',rfe),('m',model)])
          pipeline.fit(X_train, y=y_train)
          pipeline.predict(pd.DataFrame(X_test.iloc[1].values.reshape(1,-1)))
          C:\Users\jmspa\anaconda3\lib\site-packages\sklearn\base.py:450: UserWarning: X does not have valid fe
          ature names, but RFE was fitted with feature names
            warnings.warn(
Out[103]: array([1574.69506925])
```

Pengujian kedua juga memberi hasil yang cukup memuaskan dengan apartemen berharga 1500 diprediksi sebagai 1574.69506925

Apartemen pertama dengan harga 788, diprediksi oleh model dengan harga 780. Apartemen kedua dengan harga 1500, diprediksi oleh model dengan harga 1574. Pengujian ini menunjukkan hasil prediksi model yang cukup memuaskan.

"Di hadapan TUHAN yang hidup, saya menegaskan bahwa saya tidak memberikan maupun menerima bantuan apapun—baik lisan, tulisan, maupun elektronik—di dalam ujian ini selain daripada apa yang telah diizinkan oleh pengajar, dan tidak akan menyebarkan baik soal maupun jawaban ujian kepada pihak lain"

James Patrick Oentoro