



Technical Document of MariBisnis

Microsoft Capstone Project

NOV 2021

MariBisnis

ISSUED BY

MICROSOFT - MariBelajar

REPRESENTATIVE

SALSABILA ZAHIRAH PRANIDA

Microsoft Data and Artificial Intelligence

DAI 003



Outlines

[Introduction](#)

[Dataset Guidelines](#)

[Resources](#)

[Membuat Machine Learning Resources](#)

[Membuat Compute Resources](#)

[Membuat Dataset Resources](#)

[Pipeline](#)

[Membuat Pipeline](#)

[Membuat Training House Pricing](#)

[Membuat Training House Pricing - real time inference](#)

[Percobaan Endpoints](#)

[Evaluasi](#)

[Power BI Service](#)

[Mengapa Power BI Service?](#)

[Pengimporan Data MariBisnis](#)

[Visualisasi House Condition](#)

[Visualisasi House Pricing Correlated](#)

[References](#)



Introduction

MariBisnis ingin mengetahui suatu bisnis tren dalam penjualan rumah yang terjadi di Seattle, America. Data ini mengandung penjualan rumah pada bulan May 2014 dan May 2015. Terdapat dua goals dari hal ini:

1. Memprediksi harga dari sebuah unit rumah yang terjual.
2. Memetakan sebaran dari data yang ada.



Dataset Guidelines

Berikut adalah informasi mengenai dataset MariBisnis:

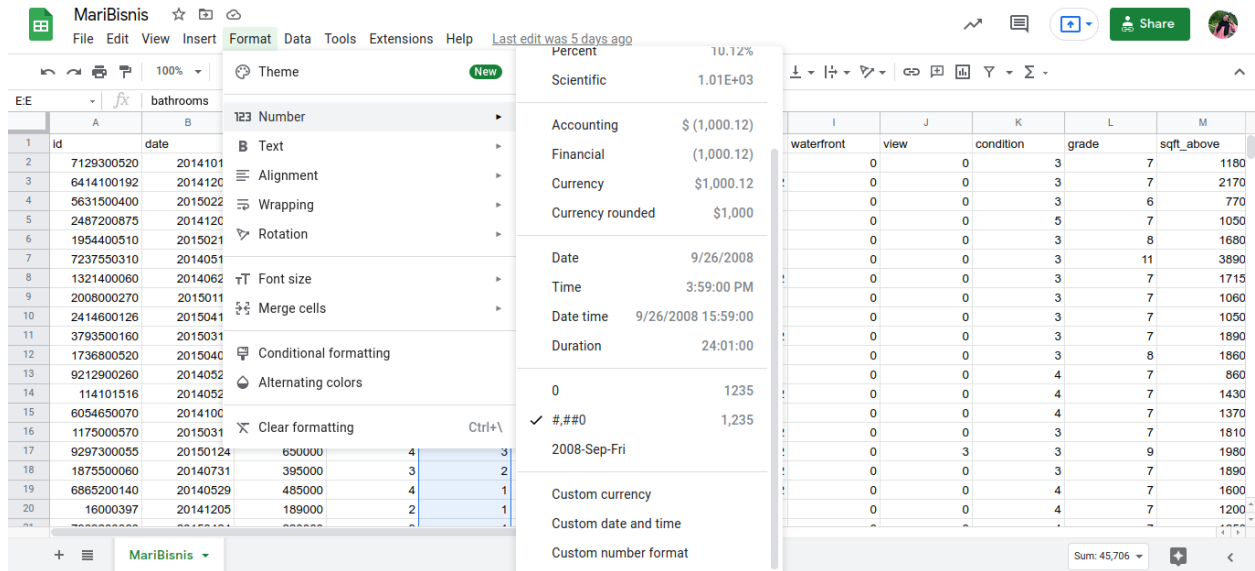
Version	: v1.0
Properties	: Tabular
Size	: 2,065 Mb
Total rows	: 21,613 rows
Features	: 21 features
Link Download	: Click here to download

Tabel 1. MariBisnis Dataset

Column name	Description
id	: A notation for a house
date	: Date house was sold
price	: Price is the prediction target
bedrooms	: Number of Bedrooms/House
bathrooms	: Number of bathrooms/bedrooms

sqft_living	: Square footage of the home
sqft_lot	: Square footage of the lot
floors	: Total floors (levels) in house
waterfront	: House which has a view to a waterfront
view	: Has been viewed
condition	: How good the condition is Overall
grade	: The overall grade is given to the housing unit, based on the King County grading system
sqft_above	: square footage of house apart from the basement
sqft_basement	: square footage of the basement
yr_built	: Built Year
yr_renovated	: The year when the house was renovated
zipcode	: zip code
lat	: Latitude coordinate
long	: Longitude coordinate
sqft_living15	: Living room area in 2015(implies-- some renovations) This might or might not have affected the lot size area
sqft_lot15	: Lot size area in 2015(implies-- some renovations)

Sebelum menuju langkah pembuatan *training* dan *deployment* model, dataset MariBisnis akan diolah lebih dulu di Google Spreadsheet. Pengolahan ini bertujuan untuk mengubah *value* pada kolom bathrooms yang awalnya *decimal* atau *float* menjadi *integer*. Pada Google Spreadsheet, kolom bathrooms diubah value-nya menjadi integer dengan menggunakan *format custom number*. Hal ini hanya diimplementasikan ke kolom bathrooms karena memiliki *value* desimal, yang mana akan memunculkan ambiguitas.



Gambar 1. Changing the value of bathrooms

Resources

Membuat Machine Learning Resources

Azure Machine Learning adalah platform berbasis *cloud* untuk membangun dan mengoperasikan solusi *machine learning* di Azure. Ini mencakup berbagai fitur dan kemampuan yang membantu *data scientist* menyiapkan data, melatih model, menerbitkan layanan prediktif, dan memantau penggunaannya. Salah satu fitur ini adalah antarmuka visual yang disebut *designer*, yang dapat digunakan untuk melatih, menguji, dan menerapkan model pembelajaran mesin tanpa menulis *code* apa pun.

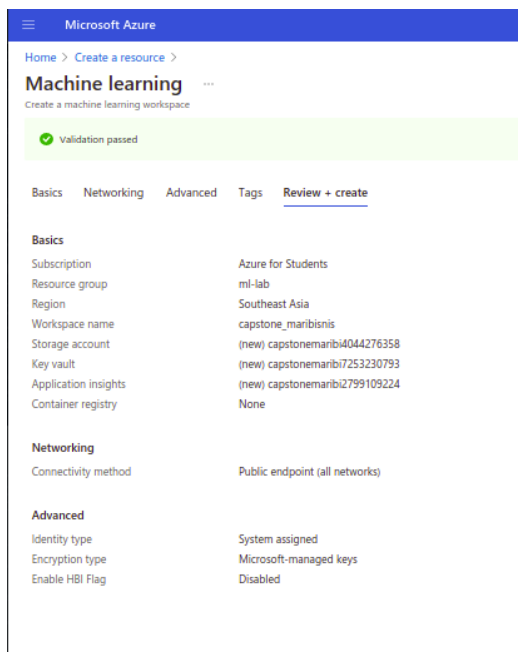
Untuk menggunakan Machine Learning Azure, diperlukan Azure *subscription*. Kemudian, *workspace* ini dapat digunakan untuk mengelola data, sumber daya komputasi, *code*, model, dan hal lain yang terkait dengan beban kerja *machine learning* Anda.

Berikut adalah langkah-langkah dalam membuat *workspace Machine Learning*:

1. Masuk ke dalam [Portal Azure](#), menggunakan kredensial Microsoft yang dimiliki.
2. Membuat *resource Machine Learning* dengan memilih + **Create a Resource** pada *home page* Azure portal.

3. Cari **Machine Learning** kemudian masukkan pengaturan sesuai arahan berikut:

Subscription	: Azure for Students
Resource group	: ml-lab
Workspace name	: capstone-maribisnis
Region	: Southeast Asia
Storage account	: capstonemaribi4044276358 (<i>Note the default new storage account that will be created for your workspace</i>)
Key vault	: capstonemaribi7253230793 (<i>Note the default new key vault that will be created for your workspace</i>)
Application insights	: capstonemaribi2799109224 (<i>Note the default new key vault that will be created for your workspace</i>)
Container registry	: None (<i>one will be created automatically the first time you deploy a model to a container</i>)



4. Pembuatan *resource* memakan waktu beberapa menit. Setelah usai, lanjutkan dengan kembali ke *home page* Azure portal.

5. Pada laman **Overview**, buka [Azure Machine Learning Studio](#) menggunakan akun Microsoft yang terdaftar Azure.

Pengelolaan *workspace* dapat menggunakan portal Azure, tetapi Azure Machine Learning Studio juga menyediakan *user interface* yang lebih terfokus untuk mengelola *resources workspace*.

Gambar 2. Machine Learning Resources

Membuat Compute Resources

Untuk melatih dan menerapkan model menggunakan desainer Machine Learning Azure, diperlukan komputasi untuk menjalankan proses pelatihan, dan untuk menguji model terlatih setelah menerapkannya.

Pada Azure Machine Learning Studio, lihat laman **Compute**. Di sinilah nantinya *compute target* akan dikelola. Terdapat empat jenis *resource compute* yang dapat dibuat:

Compute Instances	: Pengembangan <i>workstation</i> yang dapat digunakan <i>data scientist</i> untuk bekerja dengan data dan model.
Compute Clusters	: Cluster <i>virtual machine</i> yang dapat diskalakan untuk pemrosesan <i>code</i> eksperimen sesuai permintaan.
Inference Clusters	: Target penerapan untuk layanan prediktif yang menggunakan <i>trained model</i> .
Attached Compute	: Tautan ke Azure <i>compute resources</i> yang ada, seperti Virtual Machines atau cluster Azure Databricks.

Pada *Compute Cluster* tab, tambahkan *compute cluster* baru dengan pengaturan berikut. *Compute* ini akan digunakan untuk melatih model *machine learning*.

Location	: Southeast Asia (<i>Select the same as your workspace. If that location is not listed, choose the one closest to you</i>)
Virtual Machine priority	: Low Priority
Virtual Machine type	: CPU
Virtual Machine size	: Standard_E2a_v4
Compute name	: capstone
Min number of nodes	: 0
Max number of nodes	: 2
Idle before scale down	: 300 seconds
Enable SSH access	: Unselected

Compute target akan membutuhkan waktu untuk dibuat.

Attributes	Resource properties
Compute name capstone	Virtual machine size Standard_E2a_v4 (2 cores, 16 GB RAM, 50 GB disk)
Resource ID --	Processing unit CPU - Memory optimized
Compute type Machine Learning compute	OS Type Linux
Subscription ID cbd1cf3c-84fb-4788-96aa-4f37b53750ef	Virtual machine priority LowPriority
Resource group ml-lab	Minimum number of nodes 0
Workspace capstone_maribisnis	Maximum number of nodes 2
Region southeastasia	Idle seconds before scale down 300
	Virtual network/subnet --

Gambar 3. Compute Cluster Resources

Membuat Dataset Resources

Sebagai *resource* dalam pelatihan dan penerapan model nantinya di Machine Learning Azure, diperlukan akses dataset untuk sebagai sumber datanya.

Pada Azure Machine Learning Studio, lihat laman **Dataset**. Di sinilah nantinya dataset MariBisnis akan diolah menjadi lebih *compact* dan lebih bermakna.

1. Pada laman **Dataset**, pilih + **Create dataset** dan **import from local files** untuk menambahkan dataset MariBisnis.
2. Pada **Basic info**:
Name : MariBisnis
Dataset type : Tabular
Description : Maribisnis dataset for the capstone project of Microsoft Data and AI
3. Pada **Datastore and file selection**:
Select or create a datastore : workspaceblobstore
Select files for your dataset : Upload files of MariBisnis dataset

Skip data validation : Unchecked

4. Pada **Settings and preview**:

File format : Delimited

Delimiter : Comma

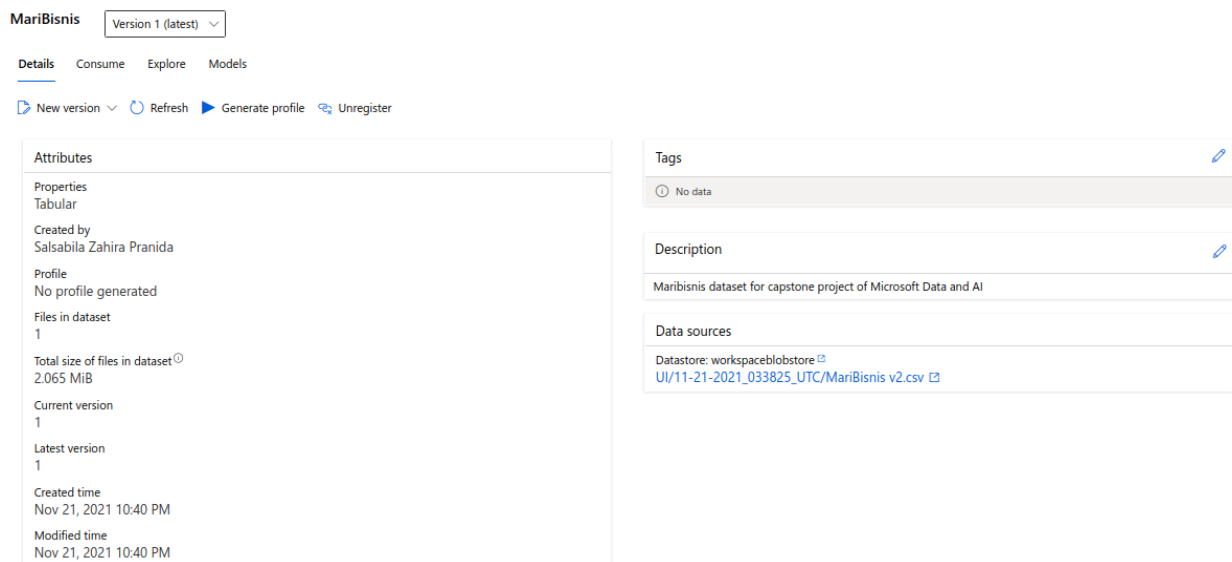
Encoding : UTF-8

Column header : All files have the same headers

Skip rows : None

5. Pada **Schema**, tidak ada perubahan, sehingga dapat langsung memilih tombol **Next**.

6. Pada **Confirm details**, setelah semuanya aman, maka bisa dilanjutkan untuk **Create**.



Gambar 4. Dataset Resources

Pipeline

Untuk melatih model regresi, diperlukan kumpulan data yang menyertakan fitur historis (karakteristik entitas yang ingin diprediksi) dan nilai label yang diketahui (nilai numerik yang ingin dilatih model untuk diprediksi).

Membuat Pipeline

Untuk menggunakan **Azure Machine Learning designer**, dapat membuat alur yang akan digunakan untuk melatih model *machine learning*. Pipeline ini dimulai dengan kumpulan data yang ingin dilatih menjadi model.

1. Pada Azure Machine Learning Studio, pada laman Designer, pilih + untuk membuat pipeline baru.
2. Pada sisi kiri atas layar, klik nama *pipeline default (Pipeline-Created-on-date)* dan ubah ke **Training House Pricing**.
3. Perlu *compute target* untuk menjalankan *pipeline*. Pilih *compute cluster* yang telah dibuat sebelumnya.

Membuat Training House Pricing

Tujuan pada pipeline **Training House Pricing** adalah untuk melatih model regresi yang akan memprediksi harga sebuah unit rumah di Seattle, America berdasarkan kondisi rumahnya. Semua kebutuhan dapat dilihat pada *panel asset library*.

1. Pada panel asset library, pilih **Dataset > MariBisnis**, kemudian *drag and drop* ke dalam kanvas.
2. Selanjutnya pilih **Data Transformation > Select Columns in Dataset**, kemudian drag and drop ke dalam kanvas.
 - a. Pada bagian ini, hubungkan dataset **MariBisnis** dengan *nodes* yang ada di **Select Columns in Dataset**.
 - b. Pada bagan **Select Columns**, isikan kolom-kolom sebagai berikut dan *Save*.



Gambar 5. Select Columns in Dataset

3. Selanjutnya pilih **Data Transformation > Edit Metadata**, kemudian *drag and drop* ke dalam kanvas.
 - a. Pada bagian ini, hubungkan **Select Columns in Dataset** dengan *nodes* yang ada di **Edit Metadata**.
 - b. Pada **Column** , isikan dengan kolom waterfront dan zipcode.
 - c. Pada **Categorical**, ubahlah menjadi Categorical.

Column ⓘ *

Column names: waterfront,zipcode

Data type ⓘ *

Unchanged

Categorical ⓘ *

Categorical

Fields ⓘ *

Unchanged

New column names ⓘ

Gambar 6. Edit Metadata

4. Selanjutnya pilih **Data Transformation > Clean Missing Data**, kemudian *drag and drop* ke dalam kanvas.
 - a. Pada bagian ini, hubungkan **Edit Metadata** dengan *nodes* yang ada di **Clean Missing Data**.
 - b. **Columns to be cleaned** , isikan dengan All columns.
 - c. **Minimum missing value ratio**: 0.0
 - d. **Maximum missing value ratio**: 1.0
 - e. **Cleaning mode**: Remove entire row. Bahwa setiap baris yang memiliki nilai kosong atau NaN akan dihapus secara otomatis pada step **Clean Missing Data**.

Columns to be cleaned ⓘ *

All columns

Minimum missing value ratio ⓘ *

0.0

Maximum missing value ratio ⓘ *

1.0

Cleaning mode ⓘ *

Remove entire row

Gambar 6. Clean Missing Data

5. Selanjutnya pilih **Data Transformation > Split Data**, kemudian *drag and drop* ke dalam kanvas.
 - a. Pada bagian ini, hubungkan **Clean Missing Data** dengan *nodes* yang ada di **Split Data**.
 - b. Pada **Splitting mode**, pilih Split Rows, karena akan membagi barisnya menjadi *training* dan *testing* dataset.
 - c. Pada **Fraction**, masukkan 0.75. Hal ini diartikan sebagai 75% dari dataset MariBisnis akan dialokasikan sebagai *training* data, dan sisanya, sebanyak 25% akan dialokasikan sebagai *testing* data.
 - d. Pada **Randomized split**, pilih True. Hal ini diartikan bahwa pemilihan dataset *training* maupun *testing* semuanya dilakukan secara teracak.
 - e. Pada **Random seed**, masukkan 123 sebagai *seed* acaknya.
 - f. Pada **Stratified split**, isikan False. Hal ini guna menghindari pemisahan baris secara berkelompok menggunakan kolom bertingkat.

Splitting mode ⓘ *

Split Rows

Fraction of rows in the first output dataset ⓘ *

0.75

Randomized split ⓘ *

True

Random seed ⓘ *

123

Stratified split ⓘ *

False

Gambar 7. Clean Missing Data

6. Selanjutnya pilih **Model Training > Train Model**, kemudian *drag and drop* ke dalam kanvas.
 - a. Pada bagian ini, hubungkan **Train Model** dengan *nodes* kanan yang ada di **Train Model**.
 - b. Pada **Label column**, isikan *price* sebagai target prediksi.

Label column ⓘ *

Column names: price

Model explanations ⓘ

False

Gambar 8. Train Model

7. Selanjutnya pilih **Machine Learning Algorithms > Decision Forest Regression**, kemudian *drag* and *drop* ke dalam kanvas.

- Pada bagian ini, hubungkan **Decision Forest Regression** dengan *nodes* kiri yang ada di **Train Model**.
- Tidak diperlukan untuk mengubah *default setting* yang ada pada algoritimanya.

Trainer mode : SingleParameter

Number of decision trees : 8

Max depth of the decision trees : 32

Min number of samples per leaf node : 1

Resampling method : Bagging Resampling

Create trainer mode ⓘ *

SingleParameter

Number of decision trees ⓘ *

8

Maximum depth of the decision trees ⓘ *

32

Minimum number of samples per leaf node ⓘ *

1

Resampling method ⓘ *

Bagging Resampling

Gambar 9. Decision Forest Regression

8. Selanjutnya pilih **Model Scoring and Evaluation > Score Model**, kemudian *drag* and *drop* ke dalam kanvas.

- a. Pada bagian ini, hubungkan **Split Data** dengan *nodes* kanan yang ada di **Score Model** dan hubungkan **Train Model** dengan *nodes* kiri yang ada di **Score Model**.
- b. Pada **Append score**, pilih True.

Score Model

Append score columns to output ⓘ *

True

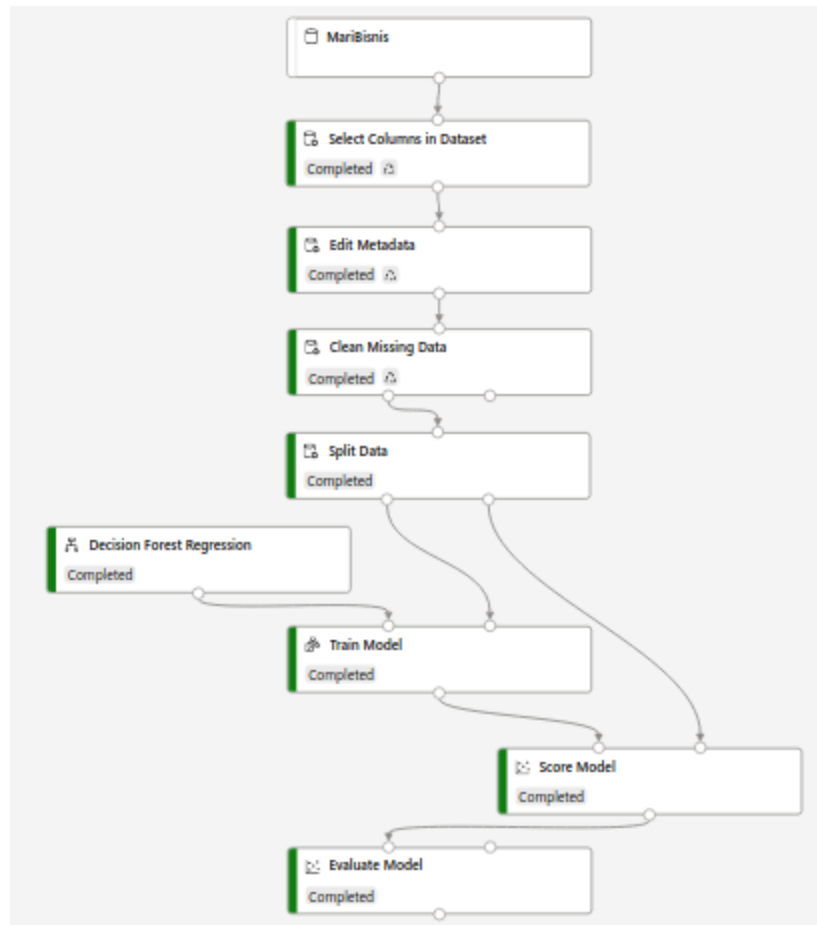
Gambar 10. Score Model

9. Selanjutnya pilih **Model Scoring and Evaluation > Evaluate Model**, kemudian *drag* and *drop* ke dalam kanvas.
 - a. Pada bagian ini, hubungkan **Score Model** dengan *nodes* kiri yang ada di **Evaluate Model**.
 - b. Pada Evaluate Model, tidak ada parameter yang perlu dimasukkan.
10. Setelah semua modul sudah berada dan terhubung di dalam kanvas, maka dapat dilakukan **Submit** dan menunggu pipeline usai *running*.
11. Setelah pipeline telah usai dibuat, pada Evaluate Model akan nampak lima matriks evaluasi sebagai berikut:

Tabel 2. Matriks Evaluasi Training

Mean Absolute Error (MAE)	:	87034.878769
Root Mean Squared Error (RMSE)	:	148735.617857
Relative Squared Error (RSE)	:	0.17349 - 17,34%
Relative Absolute Error (RAE)	:	0.375667 - 37,57%
Coefficient of Determination (R^2)	:	0.82651

Bentuk final dari Training House Pricing pipeline.



Gambar 11. Final Pipeline Training House Pricing

Membuat Training House Pricing - real time inference

Tujuan pada pipeline **Training House Pricing - real time inference** adalah untuk melatih model regresi yang akan memprediksi harga sebuah unit rumah di Seattle, America berdasarkan kondisi rumahnya. Semua kebutuhan dapat dilihat pada *panel asset library*.

1. Kanvas pada **Training House Pricing - real time inference** men-generate pipeline dari **Training House Pricing** yang telah dibuat sebelumnya.
2. Modul **Clean Missing Data** pada **Training House Pricing** diubah menjadi **TD-Training_House_Pricing-Clean_Missing_Data-Cleaning_transformation-f7a4384a** dengan **Apply Transformation** dan dihubungkan ke **Score Model**.
3. Modul **Train Model** dan **Decision Forest Regression** ditransformasikan menjadi **MD-Training_House_Pricing-Train_Model-Trained_model-e4d0e84f** dan dihubungkan ke **Score Model**.

4. Pada *panel asset library*, pilih **Web Service > Web Service Input**, kemudian *drag and drop* ke dalam kanvas. Hubungkan modul dengan *nodes* yang ada pada **Select Columns in Dataset**.
5. Pada *panel asset library*, pilih **Python Language > Execute Python Script**, kemudian *drag and drop* ke dalam kanvas. Hubungkan modul dengan *nodes* yang ada pada **Score Model**.
 - a. Pada modul **Execute Python Script**, ada penambahan *python script* sebagai berikut:

```
import pandas as pd

def azureml_main(dataframe1 = None, dataframe2 = None):

    scored_results = dataframe1[['Scored Labels']]
    scored_results.rename(columns={'Scored
Labels': 'predicted_price'},
                           inplace=True)
    return scored_results
```

6. Modul **Score Model** mengeluarkan dua *nodes* yang mengarah pada **Evaluate Model** dan **Execute Python Script**.
7. Modul **Evaluate Model** disajikan lima matriks evaluasi yang sama seperti pada pipeline **Training House Pricing**.

Tabel 3. Matriks Evaluasi Training real time inference

Mean Absolute Error (MAE)	52340.56594
Root Mean Squared Error (RMSE)	98575.012936
Relative Squared Error (RSE)	0.072098 - 7,21%
Relative Absolute Error (RAE)	0.223733 - 22,37%
Coefficient of Determination (R^2)	0.927902

8. Pada *panel asset library*, pilih **Web Service > Web Service Output**, kemudian *drag and drop* ke dalam kanvas. Hubungkan modul dengan *nodes* yang ada pada **Execute Python Script**.
9. Setelah pipeline **Training House Pricing - real time inference** selesai, maka dapat di **Submit**.
10. Setelah **Submit** berjalan dengan sukses, maka pipeline dapat diterapkan atau di **Deploy**.

11. Pada sesi *deployment*, terapkan *endpoint real time* baru menggunakan hal berikut:

Name	: predict-auto-price
Description	: This deployment for Training House Pricing
Compute type	: Azure Container Instance

12. Tunggu hingga layanan selesai, hal ini dapat memakan waktu beberapa menit.

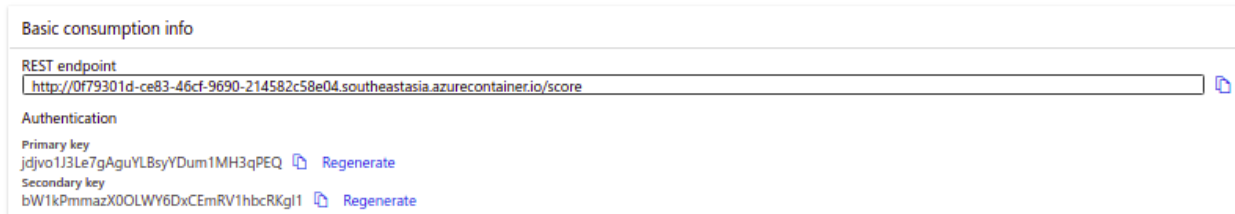
Berikut adalah bentuk final pipeline **Training House Pricing - real time inference**.



Gambar 12. Final Pipeline Training House Pricing - real time inference

Percobaan Endpoints

Setelah model dibuat dan diuji untuk inferensi *real-time*, selanjutnya adalah publikasi sebagai layanan untuk digunakan. Pada Endpoint **predict-auto-price**, buka **Consume** tab dan perhatikan informasi mengenai **REST endpoint** dan **Primary Key**.



Basic consumption info

REST endpoint

Authentication

Primary key
jdjvo1j3Le7gAguYLBsyYDum1MH3qPEQ [Regenerate](#)

Secondary key
bW1kPmmazX0OLWY6DxCEmRV1hbcRKgl1 [Regenerate](#)

Gambar 13. Consume information

Percobaan dapat dilakukan juga dengan cara membuka tab **Test**, tanpa perlu memasukkan **REST endpoint** dan **Primary Key** sebagai input. Dengan mudah dapat memasukkan input sesuai dengan dataset MariBisnis, tentunya dengan input yang berbeda. Hal ini dibantu oleh **Web Service Input** yang ada pada pipeline **Training House Pricing - real time inference**. **Web Service Output** juga berperan untuk menampilkan output dari input yang dimasukkan sebelumnya.



Test result

[parsed](#) [raw](#)

WebServiceOutput0

key	value
predicted_price	273744.157795367

Gambar 14. Web Service Output Predict



Evaluasi

Ketika merancang pipeline **Training House Pricing** dan **Training House Pricing - real time inference**, tidak lupa pada keduanya ditambahkan langkah **Evaluate Model**. Langkah ini memberikan informasi mengenai matriks evaluasi yang berjumlah lima. Berikut adalah penjelasannya:

- | | | |
|---|---|---|
| Mean Absolute Error
MAE | : | Perbedaan rata-rata antara nilai prediksi dan nilai sebenarnya. Nilai ini didasarkan pada unit yang sama dengan label, dalam hal ini dolar. Semakin rendah nilai ini, semakin baik model memprediksi. |
| Root Mean Squared Error
RMSE | : | Akar kuadrat dari perbedaan kuadrat rata-rata antara nilai yang diprediksi dan benar. Hasilnya adalah metrik berdasarkan unit yang sama dengan label (dolar). Jika dibandingkan dengan MAE (di atas), perbedaan yang lebih besar menunjukkan varians yang lebih besar dalam kesalahan individu (misalnya, dengan beberapa kesalahan sangat kecil, sementara yang lain besar). |
| Relative Squared Error
RSE | : | Metrik relatif antara 0 dan 1 berdasarkan kuadrat selisih antara nilai prediksi dan nilai sebenarnya. Semakin mendekati 0 metrik ini, semakin baik performa model. Karena metrik ini relatif, metrik ini dapat digunakan untuk membandingkan model yang labelnya berada dalam unit yang berbeda. |
| Relative Absolute Error
RAE | : | Metrik relatif antara 0 dan 1 berdasarkan perbedaan absolut antara nilai prediksi dan nilai sebenarnya. Semakin mendekati 0 metrik ini, semakin baik performa model. Seperti RSE, metrik ini dapat digunakan untuk membandingkan model yang labelnya berada di unit yang berbeda. |
| Koefisien Determinasi
R² | : | Metrik ini lebih sering disebut sebagai R-Squared , dan merangkum berapa banyak varians antara nilai prediksi dan nilai sebenarnya yang dijelaskan oleh model. Semakin mendekati 1 nilai ini, semakin baik kinerja model. |

Terdapat perbedaan dalam evaluasi keduanya, yakni sebelum prediksi dilakukan yaitu saat model hanya di-*training*, dan ketika prediksi sudah dilakukan yaitu saat model telah usai di-*training* dan siap menjadi *predictive service*.

	Training House Pricing	Training House Pricing - real time inference
MAE	87034.878769	52340.56594
RMSE	148735.617857	98575.012936
RSE	0.17349 - 17,34%	0.072098 - 7,21%
RAE	0.375667 - 37.57%	0.223733 - 22,37%
R²	0.82651	0.927902

Dari hasil matriks evaluasi yang dipaparkan di atas, dapat dilihat bahwa hasil ketika pipeline telah mengalami **Real-time inference** lebih baik. Pada **MAE**, absolut error yang terjadi semakin mengecil dari 87,034 menjadi 52,340. Kemudian pada **RMSE**, individual error pun juga semakin mengecil dari 148,735 menjadi 98,575. Selanjutnya, pada matriks **RSE** dan **RAE**, sama-sama mengalami penurunan yang signifikan dan semakin mendekati angka 0. Terakhir, pada **R²** nilai yang didapatkan menjadi lebih baik dan semakin mendekati angka 1.



Power BI Service

Power BI adalah kumpulan layanan perangkat lunak, aplikasi, dan konektor yang bekerja bersama untuk membantu, membuat, berbagi, dan menggunakan wawasan bisnis dengan cara yang paling efektif untuk layanan bisnis. [Microsoft Power BI service](#), terkadang disebut sebagai Power BI online, adalah bagian SaaS (Software as a Service) dari Power BI. Di layanan Power BI, dasbor membantu untuk memantau perkembangan bisnis. Dasbor menampilkan *tiles*, yang dapat dipilih untuk membuka laporan guna menjelajahi lebih lanjut. Dasbor dan laporan terhubung ke kumpulan data yang menyatukan semua data yang relevan di satu tempat.

Mengapa Power BI Service?

Karena adanya keterbatasan dalam *operating system* yang digunakan, berupa Linux Elementary OS based on Ubuntu 18.04 LTS (Bionic Beaver), maka digunakanlah layanan Power BI Service yang dapat dijalankan secara *online* lewat browser.

Pengimporan Data MariBisnis

Tidak jauh berbeda dengan Power BI Desktop, pengambilan data pada Power BI Service cukup mudah, yakni dengan memilih **Get data** dan memilih pengimporan melalui file *.csv yang sebelumnya sudah diolah. Pengambilan file dilakukan melalui **Local File** dan memilih MariBisnis dataset, yang kemudian akan diproses dan diletakkan ke dalam My Workspace.

Setelah pengimporan file dataset selesai, Power BI Service akan men-generate sebuah **Reports** dan sebuah **Dashboard** yang nantinya akan membantu dalam visualisasi. Visualisasi nantinya akan dilakukan pada Report file, sedangkan kumpulan visualisasi yang hendak ditampilkan sebagai final akan diletakkan di Dashboard.

Visualisasi yang akan dibuat pada MariBisnis ialah mengenai kondisi rumah - House Condition, dan korelasi **price** terhadap kolom lainnya - Correlation. Sebelum mengarah pada visualisasi di Power BI, perlu diketahui mengenai korelasi **price**. Pemrosesan untuk mengetahui korelasi ini dilakukan dengan *scripting python code*.

```
[ ] data.corr()['price'].sort_values()

zipcode      -0.053203
long         0.021626
condition     0.036362
yr_built     0.054012
sqft_lot15   0.082447
sqft_lot     0.089661
yr_renovated 0.126434
floors       0.256794
waterfront   0.266369
lat          0.307003
bedrooms     0.308350
sqft_basement 0.323816
view         0.397293
bathrooms    0.525138
sqft_living15 0.585379
sqft_above   0.605567
grade        0.667434
sqft_living  0.702035
price        1.000000
Name: price, dtype: float64
```

From the results above, it can be seen that in addition to **price**, there are **sqft_living**, **grade**, **sqft_above**, **sqft_living15**, and **bathrooms** which are highly correlated with price.

Gambar 15. Scripting Python Code for Correlation

Dari *scripting python code* diatas, dapat diketahui bahwa kolom **price** sangat berkorelasi dengan **sqft_living**, **grade**, **sqft_above**, **sqft_living15**, dan **bathrooms**. Hal ini dapat dilihat dari hasil `sort_values()` yang menunjukkan kolom-kolom yang nilainya mendekati angka 1. Nilai-nilai berikut nantinya akan membantu dalam memvisualisasikan prediksi harga sebuah unit rumah yang menjadi tujuan dari MariBisnis.

Tabel 4. Nilai Korelasi Terhadap Kolom price

	sqft_living	grade	sqft_above	sqft_living15	bathrooms
price	0.702	0.667	0.606	0.585	0.525

Visualisasi House Condition

Pada visualisasi yang pertama, bertujuan untuk mengetahui kondisi sebaran unit rumah yang terjual. Visualisasi ini menggunakan 3 (tiga) buah *chart* yang menjelaskan mengenai kondisi rumah dan grade rumah.

- Visualisasi grade unit rumah berdasarkan rerata jumlah kamar tidur dan kamar mandi; *average of bedrooms and bathrooms by grade*.

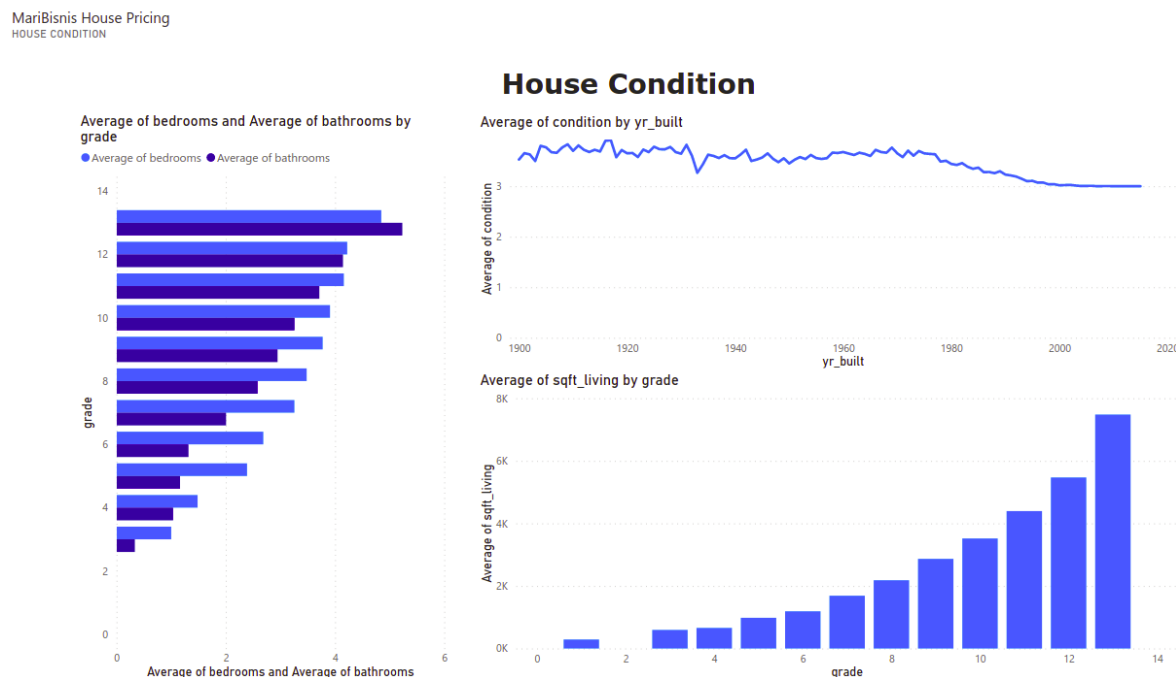
Visualisasi ini dimaksudkan untuk mengetahui sebaran kamar tidur dan kamar mandi berdasarkan grade sebuah unit rumah. Chart yang digunakan adalah **clustered bar chart**. Hasil dari analisisnya ialah semakin banyak kamar tidur dan kamar mandi pada sebuah unit rumah, maka unit rumah tersebut memiliki grade yang tinggi, berdasarkan grading system King Country.

- Visualisasi kondisi unit rumah berdasarkan tahun dibangunnya rumah; *condition by yr_built*.

Visualisasi ini dimaksudkan untuk mengetahui sebaran kondisi sebuah rumah berdasarkan tahun rumah tersebut dibangun. Chart yang digunakan adalah **line chart**. Hasil dari analisisnya ialah kondisi rumah yang dibangun pada tahun 1900-an ternyata dinilai lebih baik dibanding dengan kondisi rumah yang dibangun pada tahun 2000-an. Dapat dinyatakan bahwa, rumah lama tidak kalah bagusnya dengan unit rumah yang belum lama dibangun.

- Visualisasi grade unit rumah berdasarkan luas *living area* rumah; *sqft_living* by grade. Visualisasi ini dimaksudkan untuk mengetahui sebaran area tinggal yang dimiliki unit rumah terhadap grade. Chart yang digunakan adalah **stacked bar chart**. Hasil dari analisisnya dapat dilihat dari bar yang semakin naik, hal ini juga berbanding lurus dengan grade dengan angka yang semakin besar dan **sqft_living** yang menunjukkan luas semakin besar pula. Dapat diambil simpulan bahwa, ketika grade suatu unit rumah besar, maka area tinggal yang dimiliki rumah tersebut juga luas.

Berikut adalah visualisasi dari Report House Condition.



Gambar 16. Visualization of House Condition

Visualisasi House Pricing Correlated

Pada visualisasi yang kedua, bertujuan untuk mengetahui apa saja yang menyebabkan atau berkorelasi dengan **price** sebuah unit rumah. Visualisasi ini menggunakan 5 (lima) buah *chart* yang menjelaskan mengenai hal yang berkorelasi terhadap prediksi rumah.

- Visualisasi grade unit rumah berdasarkan rerata harga; *average of price by grade*. Visualisasi ini dimaksudkan untuk mengetahui hal yang berkorelasi dengan harga sebuah unit rumah. Chart yang digunakan adalah **stacked column chart**. Hasil dari analisisnya ialah semakin besar grade yang dimiliki sebuah unit rumah, maka hal tersebut berbanding

lurus dengan harganya. Dimaksudkan, bila sebuah rumah memiliki grade tinggi, maka harga dari rumah tersebut juga tinggi. Grade ini dinilai berdasarkan grading system King Country.

- Visualisasi bathrooms unit rumah berdasarkan rerata harga; average of price by bathrooms.

Visualisasi ini dimaksudkan untuk mengetahui hal yang berkorelasi dengan harga sebuah unit rumah. Chart yang digunakan adalah **stacked column chart**. Hasil dari analisisnya ialah semakin banyak kamar mandi yang dimiliki sebuah unit rumah, maka hal tersebut berbanding lurus dengan harganya. Dimaksudkan, bila sebuah rumah memiliki banyak kamar mandi, maka harga dari rumah tersebut juga tinggi.

- Visualisasi price unit rumah berdasarkan rerata square footage rumah; average of sqft_above by price.

Visualisasi ini dimaksudkan untuk mengetahui hal yang berkorelasi dengan harga sebuah unit rumah. Chart yang digunakan adalah **line chart**. Hasil dari analisisnya ialah semakin besar square footage semua area tinggal (tidak termasuk area basement) yang dimiliki sebuah unit rumah, maka hal tersebut berbanding lurus dengan harganya. Dimaksudkan, bila sebuah rumah memiliki square footage semua area tinggal yang luas, maka harga dari rumah tersebut juga tinggi.

- Visualisasi price unit rumah berdasarkan rerata square footage area tinggal; average of sqft_living by price.

Visualisasi ini dimaksudkan untuk mengetahui hal yang berkorelasi dengan harga sebuah unit rumah. Chart yang digunakan adalah **line chart**. Hasil dari analisisnya ialah semakin besar square footage semua area tinggal (termasuk area basement) yang dimiliki sebuah unit rumah, maka hal tersebut berbanding lurus dengan harganya. Dimaksudkan, bila sebuah rumah memiliki square footage semua area tinggal yang luas, maka harga dari rumah tersebut juga tinggi.

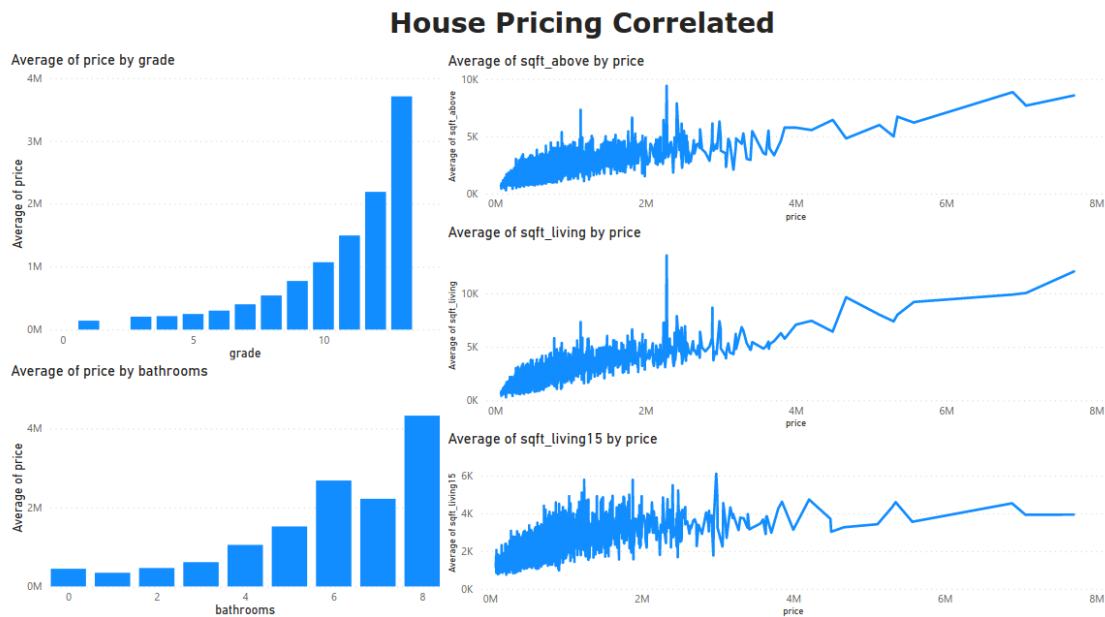
- Visualisasi price unit rumah berdasarkan rerata square footage area tinggal pada tahun 2015, baik direnovasi maupun tidak; average of sqft_living15 by price.

Visualisasi ini dimaksudkan untuk mengetahui hal yang berkorelasi dengan harga sebuah unit rumah. Chart yang digunakan adalah **line chart**. Hasil dari analisisnya ialah semakin

besar square footage semua area tinggal pada tahun 2015 (termasuk area basement, baik unit rumah mengalami renovasi ataupun tidak) yang dimiliki sebuah unit rumah, maka hal tersebut berbanding lurus dengan harganya. Dimaksudkan, bila sebuah rumah memiliki square footage semua area tinggal yang luas pada tahun 2015, maka harga dari rumah tersebut juga tinggi.

Berikut adalah visualisasi dari Report House Pricing Correlated.

MariBisnis House Pricing
CORRELATION



Gambar 17. Visualization of House Pricing Correlated

References

[Create a Regression Model with Azure Machine Learning designer](#)

[Tutorial: Get started creating in the Power BI service](#)