DAI-003

Capstone Project - Microsoft Data and Artificial Intelligence

MariBisnis

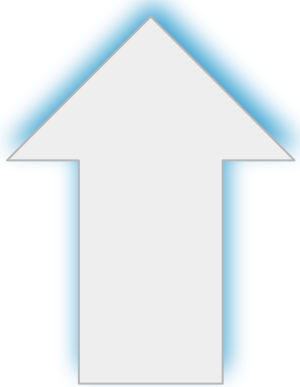
Salsabila Zahirah Pranida

MariBisnis

DELIVERABLES

MariBisnis wants to know the business trend of selling homes in Seattle, America.

The goal is that MariBisnis can **predict the price of a house and map out the distribution** of existing data.



Predict the price of a house

MariBisnis

DATASET

However...

The dataset **is not explained and trained well**. So, we need to understand the dataset. This dataset contains house sale prices, including homes sold between **May 2014 and May 2015**.

21 features that need to be explained and processed.

MariBisnis

DATASET

21 features that need to be explained and processed.

Column name	Description	Column name	Description
id	a notation for a house	date	date house was sold
price	price is prediction target	bedrooms	number of bedrooms per house
bathrooms	number of bathrooms per 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 waterfornt	view	has been viewed
condition	how good the condition is overall	grade	overall grade given to the housing unit, based on the grading system
sqft_above	square footage of house apart from basement	sqft_basement	square footage of the basemenet
yr_built	built year	yr_renovated	year when house was renovated
zipcode	zipcode	lat	latitude coordinate
long	longitude coordinate	sqft_living15	living room area in 2015 (implies-- some renovation), might or might not have affected the lotsize
sqft_lot15	lot size area in 2015 (implies-- some renovations		

MariBisnis

MariBisnis processing steps.

1

Understanding MariBisnis dataset.

2

Create compute clusters and pipeline needed.

3

Create visualization by MariBisnis's insights.

Understanding MariBisnis dataset.

MariBisnis

Dataset details.

Version	v1.0
Properties	Tabular
Size	2,065 Mb
Total rows	21,613 rows
Features	21 Features
Link download	One Drive MariBisnis Dataset

	id	date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view
0	7129300520	20141013T000000	221900.0	3	1.00	1180	5650	1.0	0	0
1	6414100192	20141209T000000	538000.0	3	2.25	2570	7242	2.0	0	0
2	5631500400	20150225T000000	180000.0	2	1.00	770	10000	1.0	0	0
3	2487200875	20141209T000000	604000.0	4	3.00	1960	5000	1.0	0	0
4	1954400510	20150218T000000	510000.0	3	2.00	1680	8080	1.0	0	0

view	condition	grade	sqft_above	sqft_basement	yr_built	yr_renovated	zipcode	lat	long	sqft_living15	sqft_lot15
0	3	7	1180	0	1955	0	98178	47.5112	-122.257	1340	5650
0	3	7	2170	400	1951	1991	98125	47.7210	-122.319	1690	7639
0	3	6	770	0	1933	0	98028	47.7379	-122.233	2720	8062
0	5	7	1050	910	1965	0	98136	47.5208	-122.393	1360	5000
0	3	8	1680	0	1987	0	98074	47.6168	-122.045	1800	7503

MariBisnis

Pre-processing dataset.

Name	Dtype	Name	Dtype	Name	Dtype
id	int	floors	int	yr_built	int
date	int	waterfront	int	yr_renovated	int
price	int	view	int	zipcode	int
bedrooms	int	condition	int	lat	float
bathrooms	float	grade	int	long	float
sqft_living	int	sqft_above	int	sqft_living15	int
sqft_lot	int	sqft_basement	int	sqft_lot15	int

Unused Features

Change into categorical features

Change dtype

MariBisnis

Final processing dataset.

Name	Dtype	Name	Dtype	Name	Dtype
price	int	condition	int	sqft_living15	int
bedrooms	int	grade	int	sqft_lot15	int
bathrooms	int	sqft_above	int		
sqft_living	int	sqft_basement	int		
sqft_lot	int	yr_built	int		
floors	int	yr_renovated	int		
waterfront	int	zipcode	int		

Unused Features

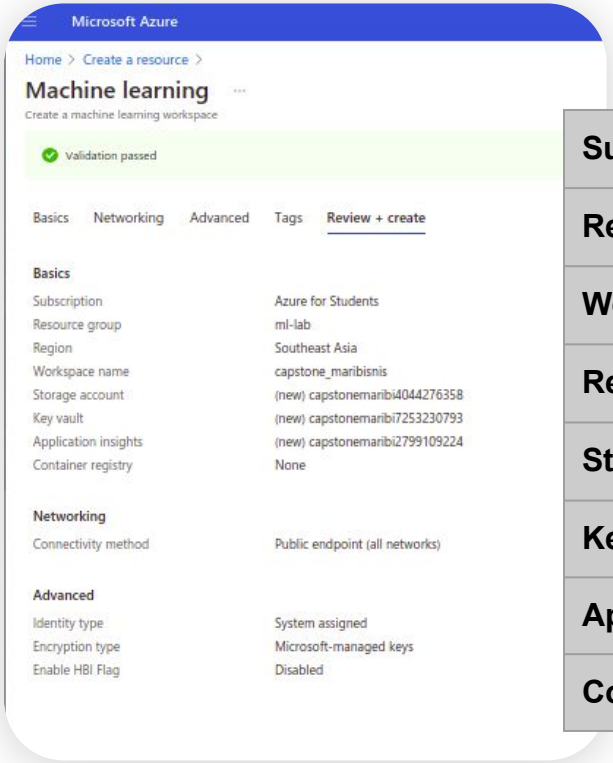
Change into categorical features

Change dtype

Create compute clusters and pipeline needed.

MariBisnis

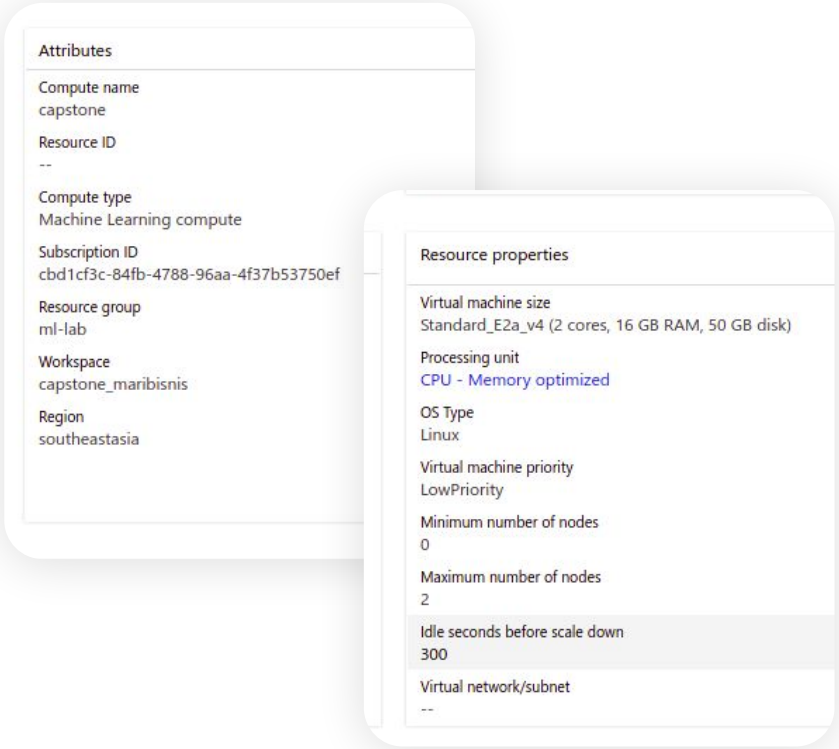
Create Machine Learning Resources.



Subscription	Azure for Students
Resource group	ml-lab
Workspace name	capstone-maribisnis
Region	Southeast Asia
Storage account	capstonemaribi4044276358
Key vault	capstonemaribi7253230793
Application insights	capstonemaribi2799109224
Container registry	None

MariBisnis

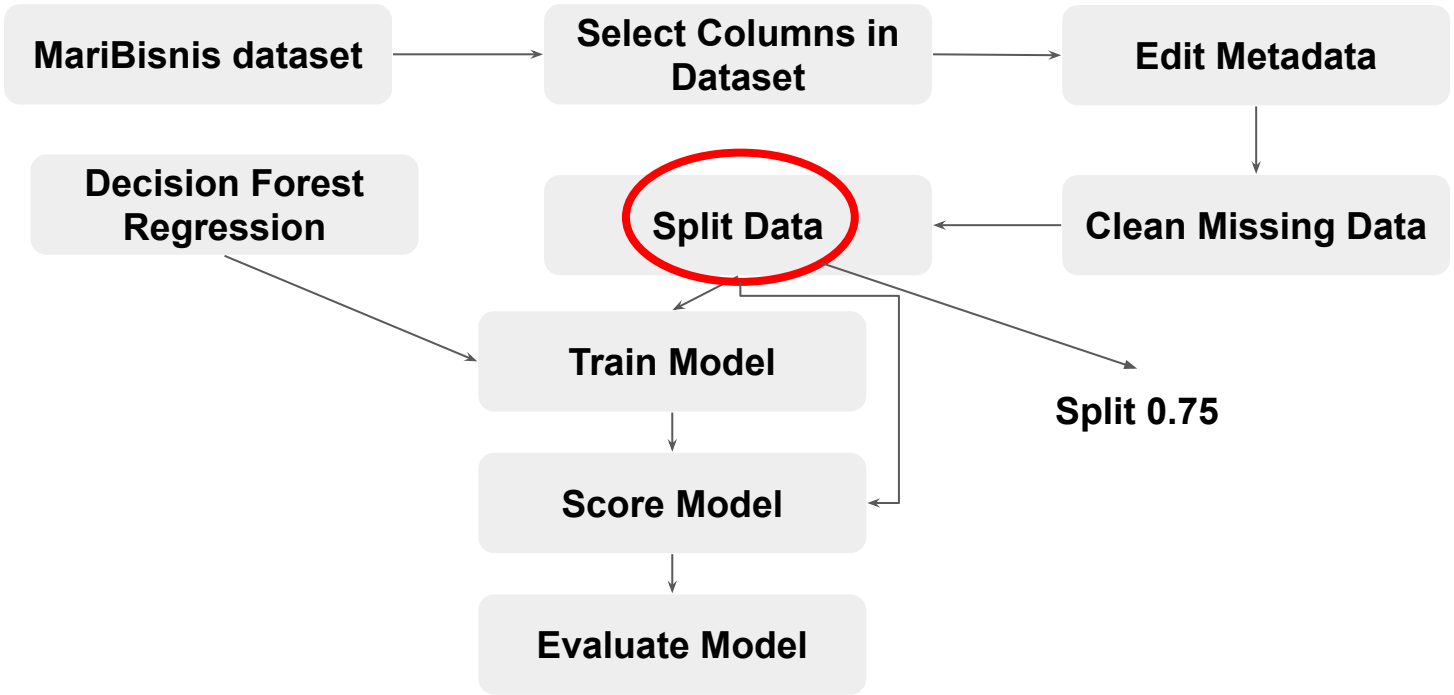
Create Compute Clusters Resources.



Location	Southeast Asia
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 seconds before scale down	300s
Enable SSH access	Unselected

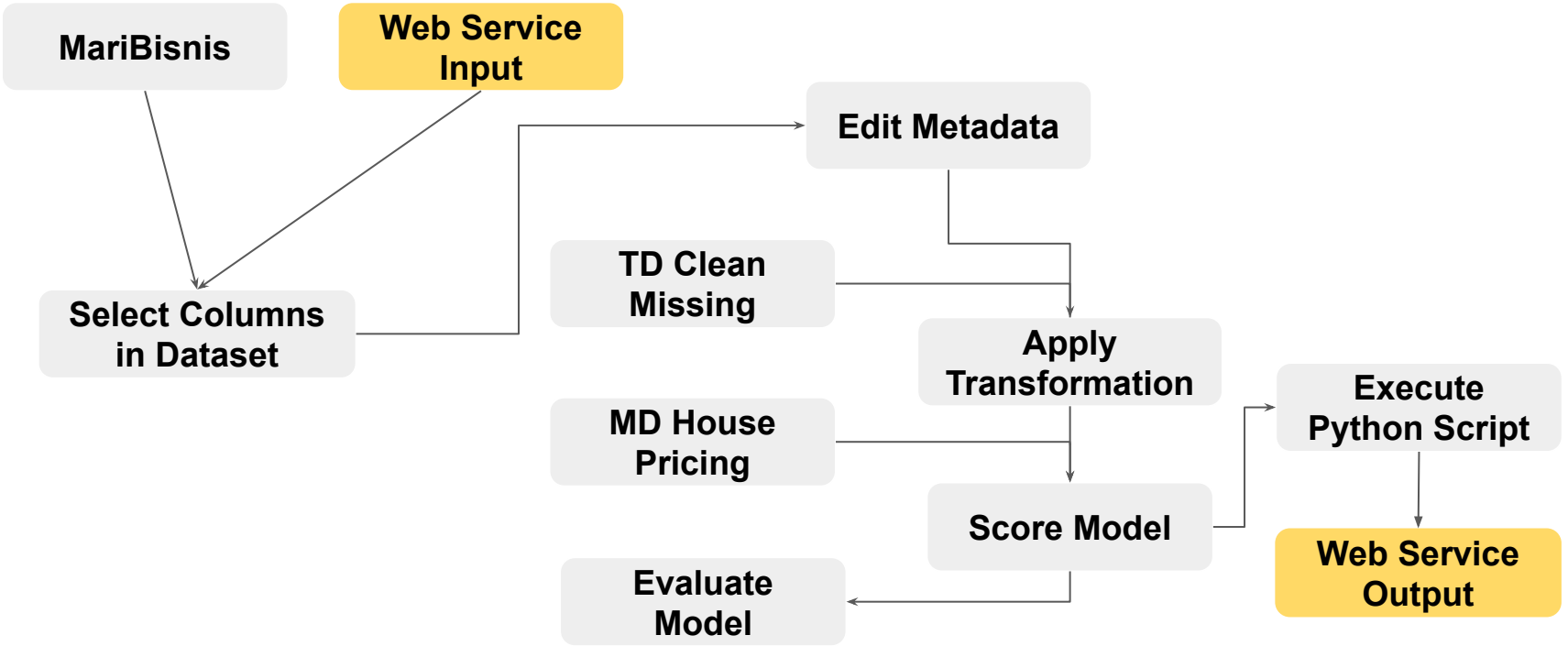
MariBisnis

Auto Training House Pricing.



MariBisnis

Auto Training House Pricing Pipeline - real time interface.



Python Script - Pipeline

```
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
```

Evaluate Model - Decision Forest Regression

MAE	52340.56594
RMSE	98575.012936
RSE	0.072098 - 7,21%
RAE	0.223733 - 22,37%
R^2	0.927902

represents the mean absolute error between the predicted result and the original result.

represents that the variation in the value produced by a forecast model is close to the variation in the observed value.

a relative metric between 0 and 1 based on the square of the differences between predicted and true values.

a relative metric between 0 and 1 based on the absolute differences between predicted and true values.

represents summarizes how much of the variance between predicted and t values is explained by the model.

Create visualization by MariBisnis's insights.

Correlation House Pricing

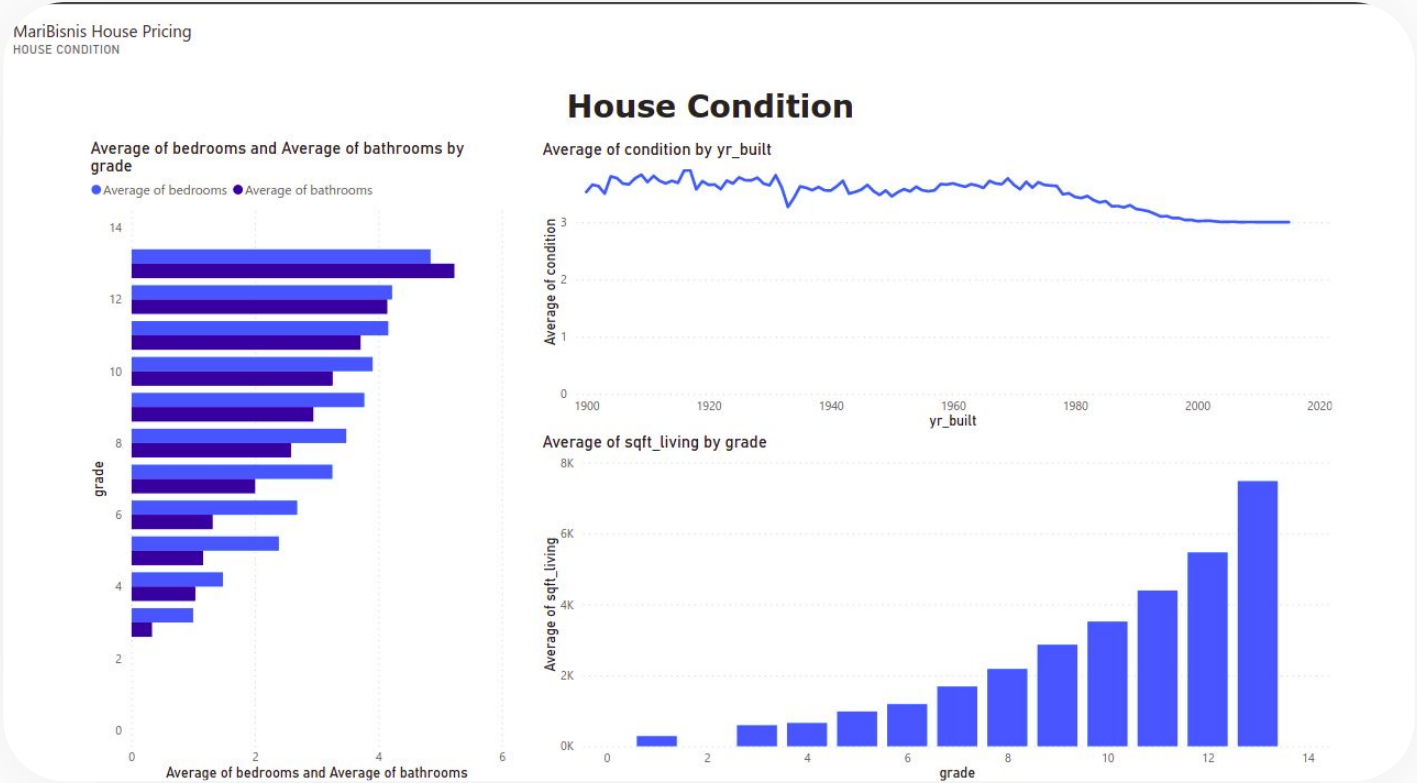
```
[ ] data.corr()['price'].sort_values()
```

```
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.

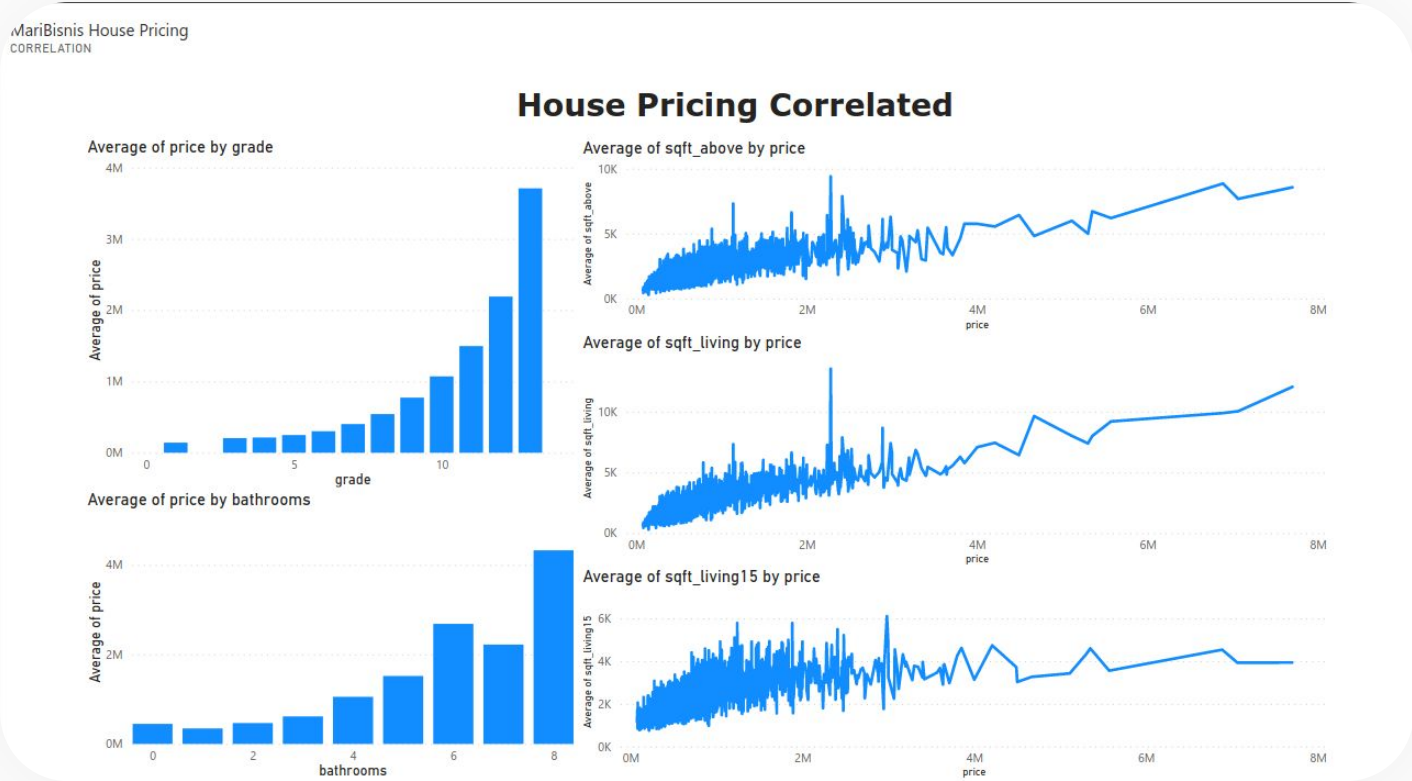
MariBisnis

House Condition Visualization



MariBisnis

House Pricing Correlated Visualization



MariBisnis

Thank you!

Authoring

Salsabila Zahirah Pranida
Microsoft Data and Artificial Intelligence
DAI - 003

Capstone Repository:
github.com/irasalsabila/microsoft-capstone-maribisnis