



TAKE-HOME MID-TERM

ISE 543, SPRING 2023



- › For this exam, you are to create a predictive model in Azure ML Studio for the attached dataset and turn in a report as specified in the following pages. You should use whichever data preparation, modeling, and model assessment techniques that were covered in this portion of the class that you believe result in the best model.
- › It may be helpful for you to visualize the dataset using Tableau (or Excel), but it is not required, and you do not need to submit any visualizations.
- › It is due at 6:00PM on Monday, April 17
- › When you are complete, save this file as a PDF and upload it to Gradescope
- › As a reminder, the work that you submit must be done individually. Unlike the homework assignments, working together is not permitted and the graders will be looking for identical solutions.



For this exam, you will use Azure ML Studio Designer to build a regression model to predict the price of a house based on its features. The dataset you will be using has been distributed with this exam and consists of the following variables:

- › House Sale Identifier
- › House Age
- › SqFt (square footage of the house)
- › Num Bathrooms
- › Num Bedrooms
- › Average Income
- › Sales Price



Describe below (in short bullet point form) the data preparation steps that you performed as well as any data quality issues that you identified and what you did about them:

- › Identified the unnecessary columns and removed them (House Sale Identifier)
- › Removed the rows with missing values in columns House Age and Average Income
- › Identified the outliers in Num Bathrooms and put a threshold of 12. Also replaced the outliers with the mean of the rest of observations
- › Applied math operations (sqrt) to correct the skewness of Sqrt, Num Bedrooms, Average Income
- › Removed the original columns of the transformed variables
- › Removed the column Num Bathrooms as it is highly correlated to Num Bedrooms (Num Bathrooms = Num Bedrooms + 1)



Summarize below the model type which yielded the best results (in terms of the model R^2) and how you configured it








- › Best performing model is Decision Forest with 8 decision trees and bagging Resampling
- › R^2 : 84.9

Summarize below the method that you used to assess the model

- › The data is split into train and test data sets in 8:2 ratio. Both the sets are normalized separately. The train data is used to train the model and the model is evaluated on the test data to see how close are the predicted values to the observed ones.

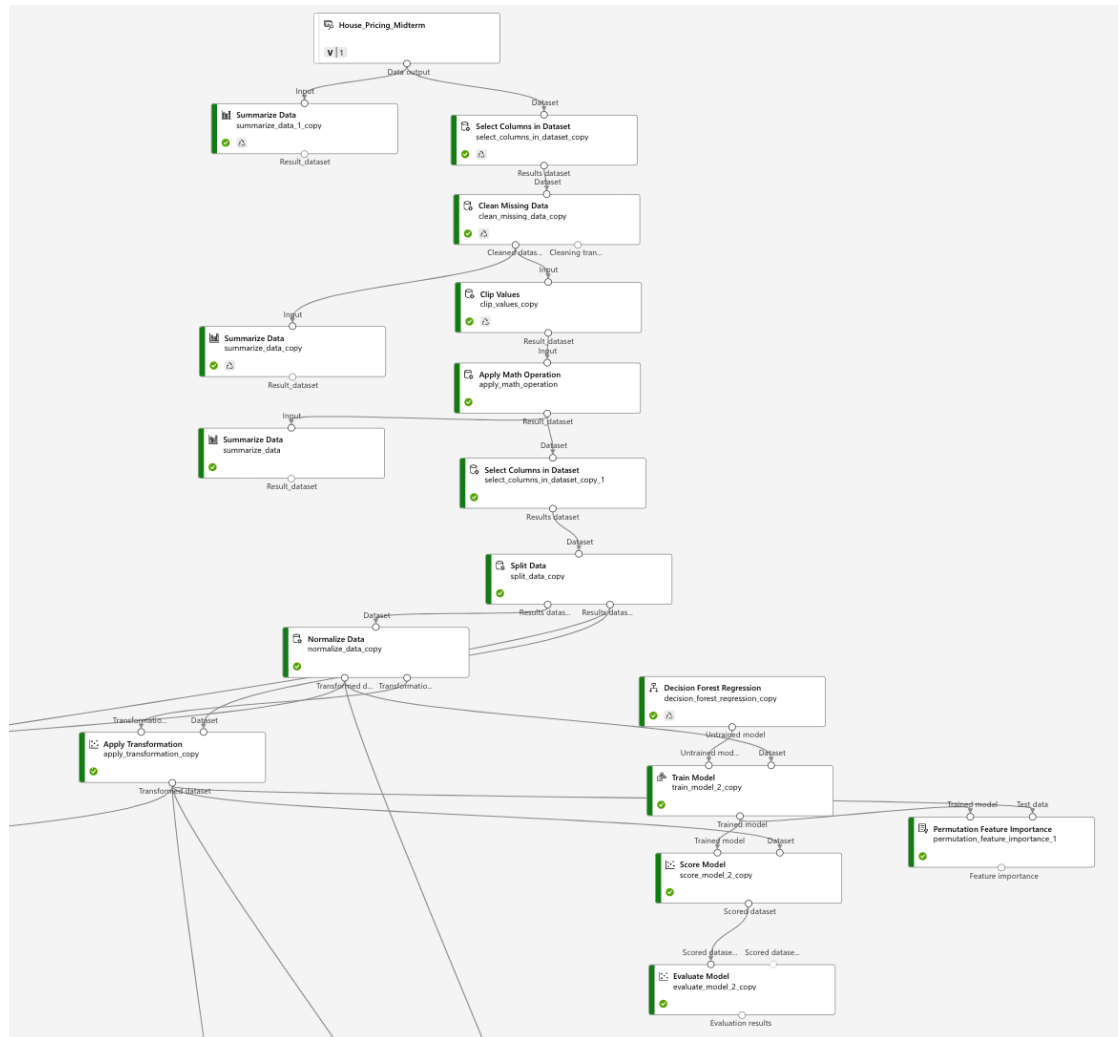


Copy below a screenshot from Azure ML studio showing the first few rows of your scored dataset (using Preview Data -> Scored Datasets) that shows a column with the Sales Price and another column with the Scored Labels:

House Age	Num Bathrooms	Sales Price	Sqrt(SqFt)	Sqrt(Num Bedrooms)	Sqrt(Average Income)	Scored Labels
						
0.776035	0.688943	260680	1.28841	0.800553	-0.687611	268147.75
0.313143	-0.378419	507120	-0.081367	-0.240147	2.178815	505453.375
-0.61264	-0.378419	322501	1.03388	-0.240147	1.122505	303789.875
0.64378	-0.378419	227293	0.111918	-0.240147	0.990904	272087.625
-0.877149	-0.9121	279511	0.585217	-0.89642	-1.20289	219802.625
-1.075532	1.756304	388161	0.666149	1.646472	2.033207	504118.75
2.16471	-0.9121	261623	-1.004021	-0.89642	0.544777	238461.375
0.379271	-0.9121	201746	-1.01158	-0.89642	1.06626	200710
0.709908	1.222623	263948	-0.467546	1.241229	-0.913069	213913.375
-0.943277	-0.9121	298192	0.373311	-0.89642	-0.899762	218954.25
-1.406168	2.289985	443265	1.905375	2.023662	1.722267	434822.75
-0.678767	1.222623	622197	-0.212118	1.241229	2.657874	615763.5
-0.282003	-1.445781	159814	-1.481767	-1.751693	-0.747286	199778.25
-0.546513	0.155262	724478	0.986061	0.313118	2.780421	639231
0.379271	0.155262	326072	0.880619	0.313118	-0.726068	297176
2.16471	0.155262	179631	-0.977622	0.313118	-0.655273	230979.875
1.767946	1.222623	355603	1.847788	1.241229	-0.747286	313508.375



Copy below a screenshot from Azure ML studio showing your final model:





Enter below the three most important variables in your final model:

- › Sqrt(Average Income)
- › Sqrt(SqFt)
- › House Age