Final Project

Azure Machine Learning

Klonicki, Steve

Deep Azure@McKesson

Dr. Zoran B. Djordjević

Problem statement

- In 2017, around 5.57 million existing homes were sold in the US
- If priced too low, the seller leaves money on the table.
- If priced too high, the house will sit on the market unsold.
- A negative perception can result when a house is on the market for a considerable amount of time, or when the price is reduced often.

Goal

- How accurately can we predict the selling price of a home before it is put onto the market?

- Build a system to predict housing prices using Azure Machine Learning.

Install/Configure/Set up

- Data set
  - https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data
  - consists of 1,460 rows of data of house sales, each with 81 attributes.

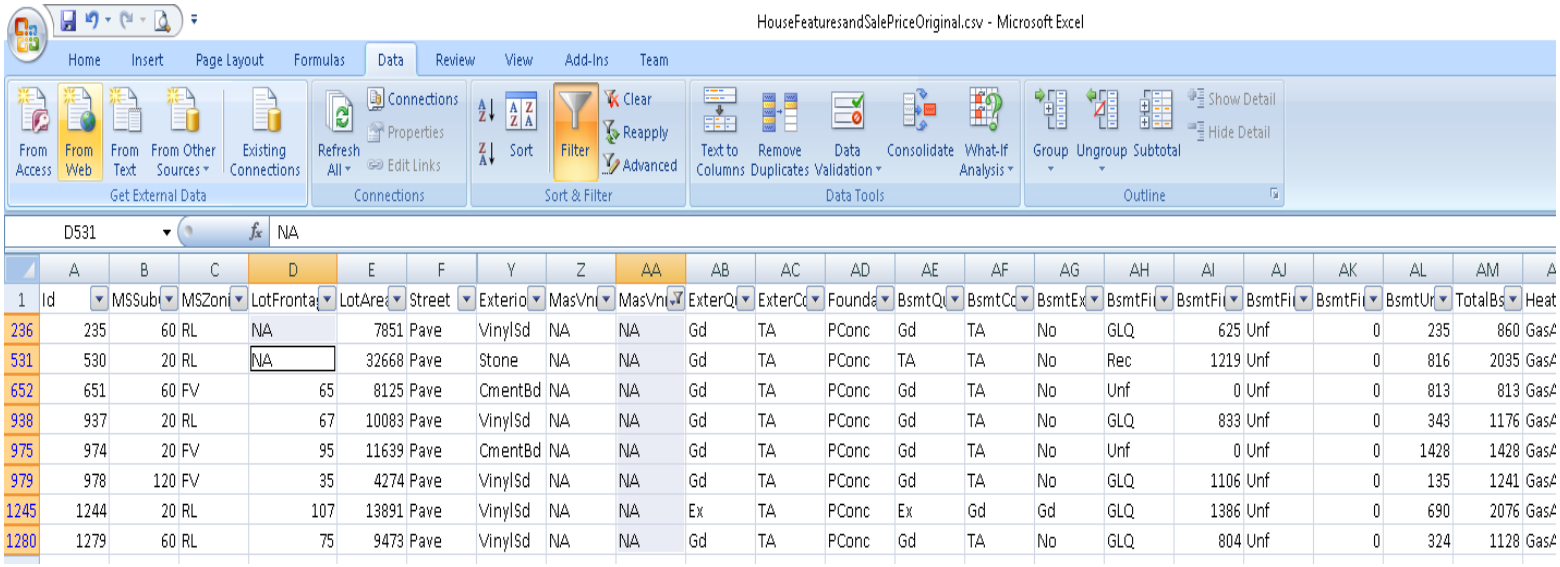
- Microsoft Azure Machine Learning Studio (https://studio.azureml.net/)

- Anaconda 5.0 distribution of Python 64 bit (https://www.anaconda.com/), which includes:
  - Python 2.7.14
  - Jupyter (visualization tool)

Install/Configure/Set up

- Data cleaning

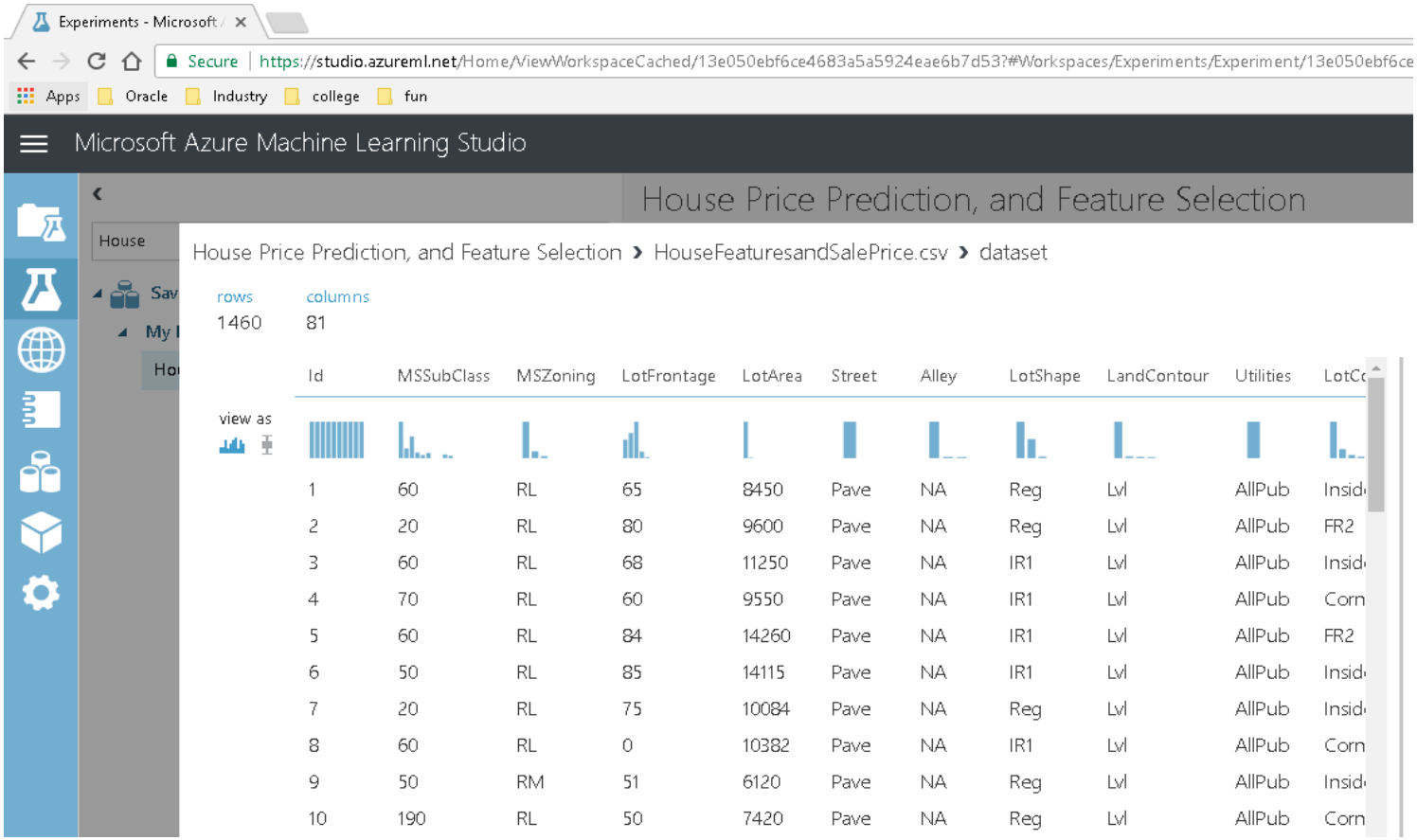
- Columns LotFrontage and MasVnrArea are numeric. The creators of the data used the value 'NA' to denote no value. In these cases, I replaced NA with '0'. Cleansing is necessary to supply correct values to the machine learning algorithms.



	A	B	C	D	E	F	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL	AM	A
1	Id	MSSub	MSZoni	LotFronta	LotArea	Street	Exterio	MasVn	MasVn	ExterQ	ExterCo	Founda	BsmtQ	BsmtCo	BsmtEx	BsmtFi	BsmtFi	BsmtFi	BsmtFi	BsmtUr	TotalBs	Heat
236	235	60	RL	NA	7851	Pave	VinylSd	NA	NA	Gd	TA	PConc	Gd	TA	No	GLQ	625	Unf	0	235	860	GasA
531	530	20	RL	NA	32668	Pave	Stone	NA	NA	Gd	TA	PConc	TA	TA	No	Rec	1219	Unf	0	816	2035	GasA
652	651	60	FV	65	8125	Pave	CmentBd	NA	NA	Gd	TA	PConc	Gd	TA	No	Unf	0	Unf	0	813	813	GasA
938	937	20	RL	67	10083	Pave	VinylSd	NA	NA	Gd	TA	PConc	Gd	TA	No	GLQ	833	Unf	0	343	1176	GasA
975	974	20	FV	95	11639	Pave	CmentBd	NA	NA	Gd	TA	PConc	Gd	TA	No	Unf	0	Unf	0	1428	1428	GasA
979	978	120	FV	35	4274	Pave	VinylSd	NA	NA	Gd	TA	PConc	Gd	TA	No	GLQ	1106	Unf	0	135	1241	GasA
1245	1244	20	RL	107	13891	Pave	VinylSd	NA	NA	Ex	TA	PConc	Ex	Gd	Gd	GLQ	1386	Unf	0	690	2076	GasA
1280	1279	60	RL	75	9473	Pave	VinylSd	NA	NA	Gd	TA	PConc	Gd	TA	No	GLQ	804	Unf	0	324	1128	GasA

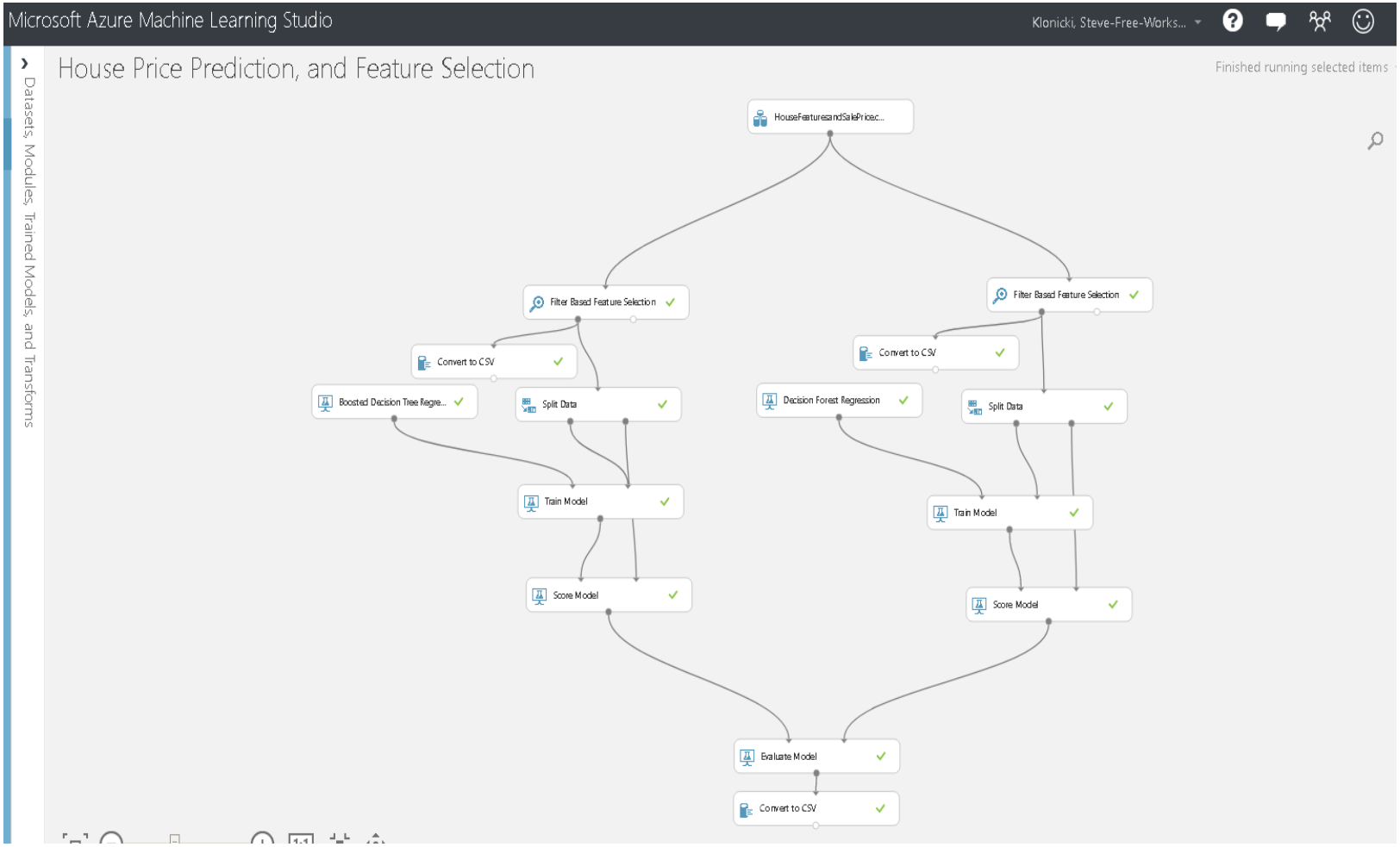
Demonstration

- Out of the box, Azure Machine Learning offers basic visualizations



Demonstration

- Machine Learning is accomplished through a series of interconnected modules



Demonstration

- Not all data features are correlated with SalesPrice

```
In [21]: frame.Street.unique()
Out[21]: array([u'Pave', u'Grvl'], dtype=object)

In [23]: frame.Street.count()
Out[23]: 1460

In [24]: frame.groupby(['Street']).size()
Out[24]: Street
         Grvl       6
         Pave    1454
         dtype: int64

In [27]: print(1454 / float(1460))
         0.995890410959
```

Demonstration

- Filter Based Feature Selection module attempts to identify the most relevant attributes.

I observed how the 2 Feature Scoring methods came to different conclusions. Listed in descending order of importance

Mutual Information	Pearson Correlation
Overall Quality	OverallQuality
Neighborhood	GrLivArea
GrLivArea	GarageCars
GarageCars	GarageArea
GarageArea	TotalBsmtSF
YearBuilt	1stFlrSF
TotalBsmtSF	FullBath
BsmtQual	TotRmsAbvGrd
ExterQual	YearBuilt
KitchenQual	YearRemodled

Demonstration

- Split the data, use 75% to train the algorithm.
- The remaining 25% was used to determine the accuracy of the model.
- Not all Regression algorithms are equal. Compare the accuracy of the 2 models against each other.
- Boosted Decision Tree Regression outperformed Decision Forest Regression. Coefficient of Determination closest to 1 wins.

In [34]:
```
experiment = ws.experiments['13e050ebf6ce4683a5a5924eae6b7d53.f-id.bd61882e87b644c29a7de2c3f741be68']
ds = experiment.get_intermediate_dataset(
    node_id='f193f8bb-2455-4dee-a6e7-76f3803cbd91-5169',
    port_name='Results dataset',
    data_type_id='GenericCSV'
)
frame = ds.to_dataframe()
```

In [35]:
```
frame
```

Out[35]:

	Negative Log Likelihood	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error	Relative Squared Error	Coefficient of Determination
0	∞	20501.911836	29997.296732	0.360987	0.141419	0.858581
1	4377.84695669526	22088.663480	32556.124367	0.388926	0.166574	0.833426

Summary

Pros

- Quickly generate an experiment that, in theory, could predict the price of a house with 85% accuracy

Cons

- Limited ways Azure Machine Learning can be used programmatically.
- Not possible to configure or execute the machine learning modules through a programming language.

YouTube URLs, GitHub URL, Last Page

- Two minute (short): https://youtu.be/kb44rs7QztQ
- 15 minutes (long): https://youtu.be/GQZ3NAmyNxA
- GitHub Repository with all artifacts: https://github.com/we814/AMLDeepDive