Report

Cloud and AI Project

Precigout Gabriel

1. Architecture

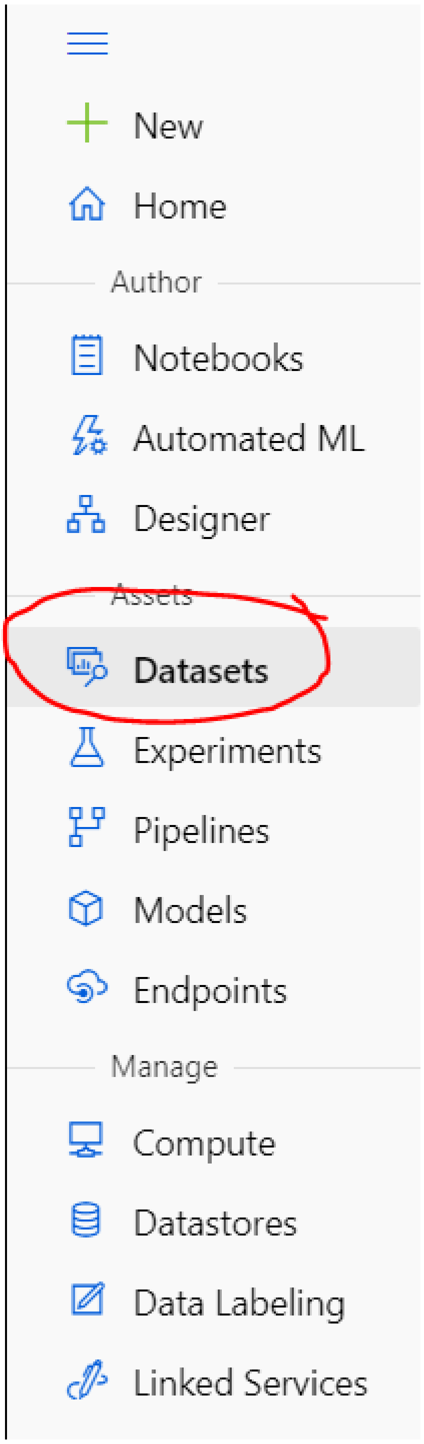
Objective: *Elaborate the architecture where you have to show the resources you are using regarding the steps of the project. This step should be based on the knowledge you gained during the tutorials and lectures. The architecture must be clear and retrace all the resources you are using. It must be included in the project report.*

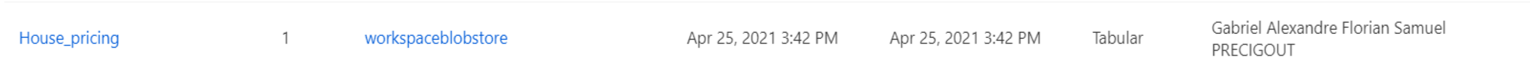
After doing and learning how to use the Microsoft Azure Machine Learning Studio with our AI and Cloud Infrastructure course, I thought about the architecture I want to implement and I want to create a pipeline to perform multiple jobs.

The first one is the preparation of the data where I will drop uncorrelated columns and create dummies for categorical data and dropping NaN values as well. I noticed some NA values that I will also drop.

The second job is the training of the model where I'll load the preprocessed data from the previous job then I'll separate features and labels, split it into train and test sets before training the model.

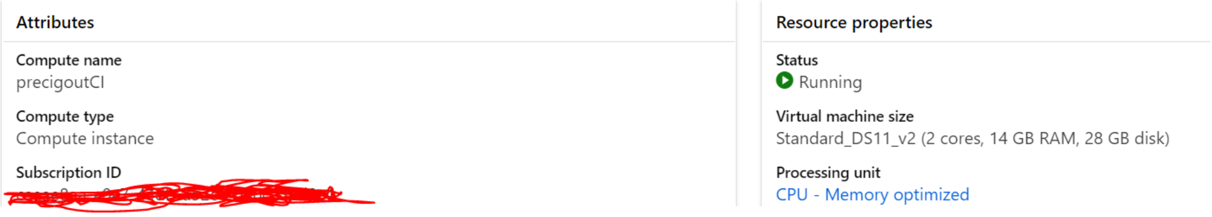
I imported my dataset on Microsoft Azure Machine Learning Studio here:





House_pricing	1	workspaceblobstore	Apr 25, 2021 3:42 PM	Apr 25, 2021 3:42 PM	Tabular	Gabriel Alexandre Florian Samuel PRECIGOUT

I also create a compute instance with the following properties:



Attributes

Compute name
precigoutCI

Compute type
Compute instance

Subscription ID
[redacted]

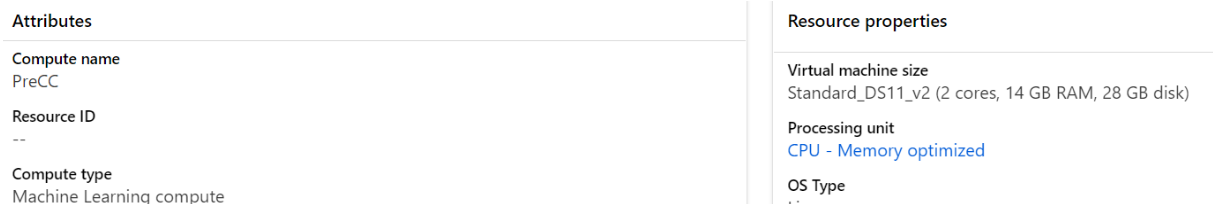
Resource properties

Status
Running

Virtual machine size
Standard_DS11_v2 (2 cores, 14 GB RAM, 28 GB disk)

Processing unit
CPU - Memory optimized

I created a Compute Instance to run my pipeline with the following parameters:



Attributes

Compute name
PreCC

Resource ID
--

Compute type
Machine Learning compute

Resource properties

Virtual machine size
Standard_DS11_v2 (2 cores, 14 GB RAM, 28 GB disk)

Processing unit
CPU - Memory optimized

OS Type

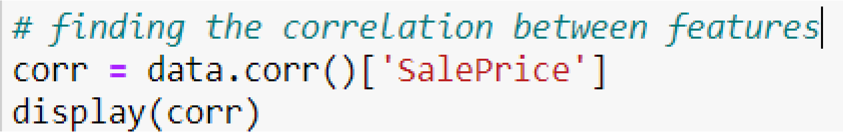
2. Data Exploration

Objective: *Put the dataset on the cloud (register it in your workspace), set up a notebook and start the adequate resources in order to make an exploratory data analysis. The exploration should be made using a notebook that use the experiment and the log capabilities in order to save the insights that you will extract in the cloud. (This notebook will be part of the rendering)*

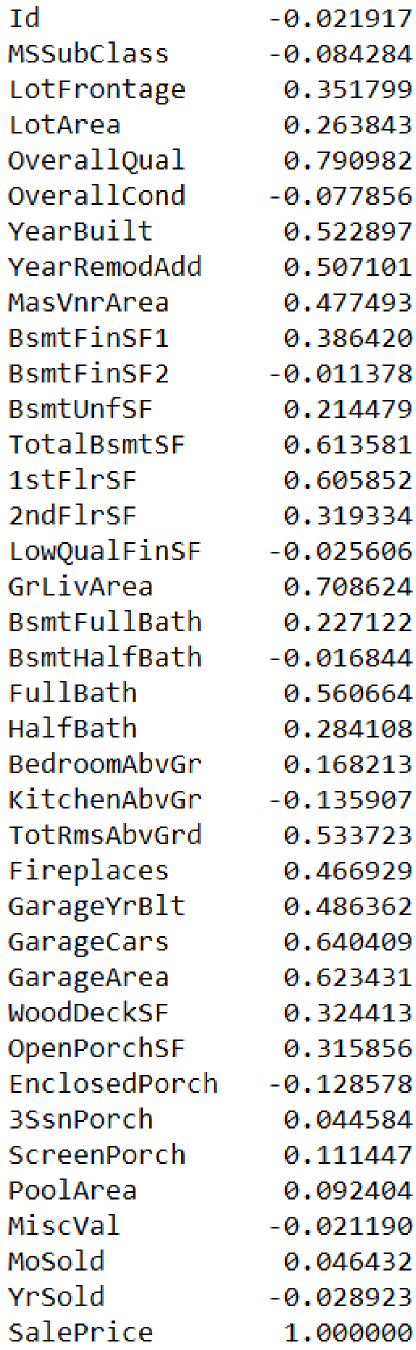
The dataset of our project is about houses and a lot of information on them as well as their selling price, the goal is to predict the price of a house according to multiple features.

As you saw above, I uploaded the dataset on my workspace and named it House_pricing then I created a Notebook called “Project_part-1” in my workspace to perform my data exploration.

First, I wanted to know the correlation between all the features and the value I want to predict which is SalePrice, so I used the corr() method and got the following:

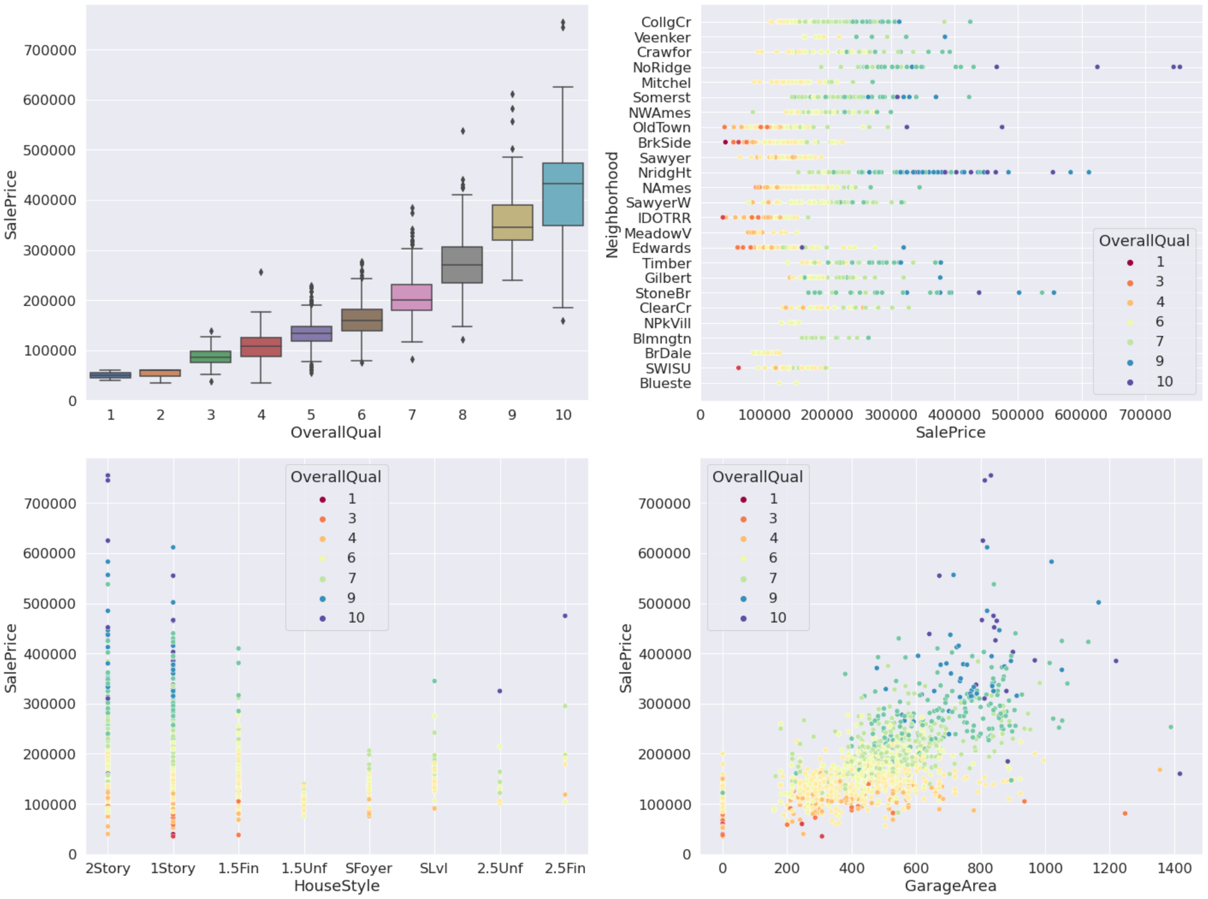


```
# finding the correlation between features
corr = data.corr()['SalePrice']
display(corr)
```



```
Id              -0.021917
MSSubClass      -0.084284
LotFrontage      0.351799
LotArea          0.263843
OverallQual      0.790982
OverallCond     -0.077856
YearBuilt        0.522897
YearRemodAdd     0.507101
MasVnrArea       0.477493
BsmtFinSF1       0.386420
BsmtFinSF2      -0.011378
BsmtUnfSF        0.214479
TotalBsmtSF      0.613581
1stFlrSF         0.605852
2ndFlrSF         0.319334
LowQualFinSF    -0.025606
GrLivArea        0.708624
BsmtFullBath     0.227122
BsmtHalfBath    -0.016844
FullBath         0.560664
HalfBath         0.284108
BedroomAbvGr     0.168213
KitchenAbvGr    -0.135907
TotRmsAbvGrd     0.533723
Fireplaces       0.466929
GarageYrBlt      0.486362
GarageCars       0.640409
GarageArea       0.623431
WoodDeckSF       0.324413
OpenPorchSF      0.315856
EnclosedPorch   -0.128578
3SsnPorch        0.044584
ScreenPorch      0.111447
PoolArea         0.092404
MiscVal         -0.021190
MoSold           0.046432
YrSold          -0.028923
SalePrice        1.000000
```

Then I decided to do some graphs using Seaborn to get some explore a bit in relation between features and prices that peaked my interest, I created 4 graphs with the sale price and the following features : Overall quality, Neighborhood type, House style, Garage area and added some color change according to the overall quality of the house.



3. Building a Machine Learning Model

Objective: *After exploring the dataset, you can start building machine learning models, elaborate a pipeline, and use the dataset you registered in your workspace, the scripts for experiments and build your machine learning solution for the problem using the corresponding azure. Do not forget to register the trained models at the end of your scripts. (The notebook will be part of your rendering).*

In order to build our Machine Learning Model, I created 2 experiment files, one for each job. The first experiment file, named “prep_housePrice.py” is going to prepare the data for the second one, I imported the dataset and replaced all the missing values and “NA” values to NaN and drop some features that were not correlated to the SalePrice such as MischFeature, Fence, PoolQC, FireplaceQu, Alley and LotFrontage, then I created some dummies resulting in a big number of columns going from 48 to 280. The new dataset is then saved as “dataHP.csv” for later use by the second experiment script.

The second experiment file, called “train_housePrice.py”, load the preprocessed dataset “dataHS.csv”, then I separate the features (X) and the value to predict (y) in order to do a train test split of the data set with 70% of the total data in the training set and 30% in the testing set. The model I chose is a decisionTreeClassifier that I’m going to use with my newly created training sets to train my model. I’ll then calculate its accuracy and save everything in an output folder.

After creating the experiments, I created the pipeline on my Compute Cluster with my two jobs that I configured and I ran everything. My accuracy is 42%, this result could be improved using a different model but my focus was to learn; be productive and efficient using Microsoft Azure Machine Learning Studio and not to have a good accuracy.

I published my pipeline on a REST endpoint:

Name: housePrice-training-pipeline

Id:

0eccd662-f0f3-4aab-ba06-6d8eb65a1c74

Endpoint: REST Endpoint

Link: https://westus.api.azureml.ms/pipelines/v1.0/subscriptions/caaae8aa-a0a7-44db-a52c-2690c23ed8c4/resourceGroups/st2aic-bd1-sg3/providers/Microsoft.MachineLearningServices/workspaces/precigout/PipelineRuns/PipelineSubmit/0eccd662-f0f3-4aab-ba06-6d8eb65a1c74

runId: 7763b2ea-7253-40a3-92de-dfa0a1df25fd

Link to Azure Machine Learning Portal: https://ml.azure.com/runs/7763b2ea-7253-40a3-92de-dfa0a1df25fd?wsid=/subscriptions/caaae8aa-a0a7-44db-a52c-

2690c23ed8c4/resourcegroups/st2aic-bd1-sg3/workspaces/precigout&tid=413600cf-bd4e-4c7c-8a61-69e73cddf731
Notebook name: "Project_part-2"

4. Batch Inferencing Pipeline

Objective: *Once you have tuned your model and registered it, create a pipeline for batch inferencing based on the best model you have obtained. (The notebook will be part of your rendering).*

I created a new Jupyter Notebook, I connected to my workspace and loaded the dataset and created multiple batches with 100 samples per batch, resulting in 100 batch files then I connected to my Compute Cluster and created my batch inferencing experiments that I will run on my pipeline. I wanted to parallelize my work but I was not able to.

Notebook name: "Project_part-3"

5. Conclusion

This project was interesting, it was fun to use Microsoft Azure instead of our own resources and computers to perform our machine learning tasks and it was faster than my old computer setup. I hope that I'll be able to use my newly acquired skills during my internship because more and more companies are turning to cloud services and a lot of them chose Microsoft Azure for their needs.