Model Selector

Harry McNinson

Outline

- Project background
- Data Sources
- Project Structure
- Regression module
- Classification module

Project Background

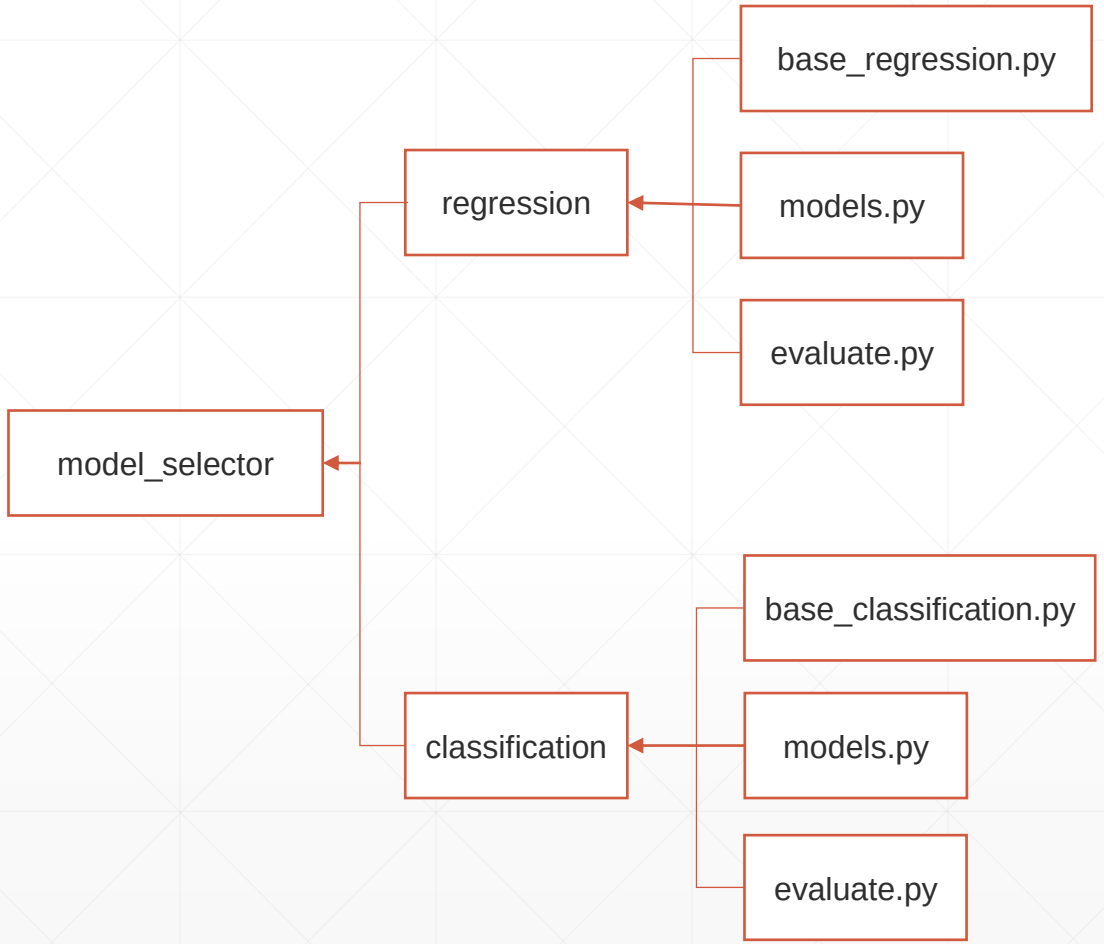
- One of the most frequently asked questions in the data science community seeks to determine which regression or classification model is best suited to be used on different datasets.
- This project aims to build a python package that will help you evaluate your regression models to select the best model for your dataset quickly and efficiently.
- Provide the file path to your dataset with some optional parameters and watch this package do the rest of the work for you.

Data Source

- Two generic datasets included in project directory
- Sales_Used_Cars.csv – To test regression models
  - 4 features
- Data_classification.csv – To test classification models
  - 11 features

Project Structure

- Model-selector is an evaluation framework for common machine learning approaches for classification and regression
- Individual modules support evaluating regression and classification models to help determine the best model for a particular dataset



Regression model selection

In [1]:
```
from model_selector.regression.evaluate import evaluate_regression
```

Testing data on Regression Models

In [8]:
```
# Provide path to dataset you want to test
file_path = "../data/Sales_Used_Cars.csv"
```

In [9]:
```
# Run dataset through regression models
result = evaluate_regression(file_path, test_size=0.2)
```

In [10]:
```
result
```

Out[10]:

	Model Name	Mean Squared Error	Mean Absolute Error	R2 Score
0	Multiple Linear	5.086716e+08	12368.960555	0.2752
1	Polynomial	4.008762e+08	9270.391649	0.4288
2	Random Forest	2.618763e+08	6995.447508	0.6268
3	Decision Tree	3.075457e+08	7509.329378	0.5618

Model with the maximum R2 score is best suited to be used for the dataset

Classification model selection

In [1]:
```
from model_selector.classification.evaluate import evaluate_classification
```

Testing data on classification models

In [16]:
```
# Provide path to dataset you want to test
file_path2 = "../data/Data_classification.csv"
```

In [13]:
```
# Run dataset through classification models
result2 = evaluate_classification(file_path2, test_size=0.25)
```

In [14]:
```
result2
```

Out[14]:

	Model Name	Accuracy Score
0	Logistic Regression	0.947368
1	Decision Tree	0.959064
2	K-Nearest Neighbors	0.947368
3	Kernel SVM	0.953216
4	Naive Bayes	0.941520
5	Random Forest	0.935673
6	Support Vector Machine	0.941520

Model with the maximum accuracy score is best suited to be used for the dataset

Challenges

- Scoping the project
- Writing unit tests
- Pip installable package

Future Works

- Add plot functionality
- Include additional ML models
- Optimize system design