MLOps ML Flow Assignment

Nitish Bhardwaj (B21AI056)

Introduction

This report presents an analysis of housing prices using two different machine learning models: Linear Regression and Random Forest Regression. The dataset used for this analysis is the Boston Housing dataset, which contains various features that influence housing prices in Boston.

Data Acquisition

The dataset was sourced from a publicly available URL. It was loaded into a Pandas DataFrame for further analysis and modeling.

Dataset Overview

The dataset consists of several features, with the target variable being medv, which represents the median value of owner-occupied homes in thousands of dollars. The features include variables such as crime rate, average number of rooms, accessibility to highways, and others.

Data Preparation

The features (X) were separated from the target variable (y). The dataset was then split into training and testing sets, with 80% of the data used for training and 20% reserved for testing.

Model Selection

Two models were chosen for the prediction task:

1. **Linear Regression**
2. **Random Forest Regressor**

These models were implemented using Scikit-learn, a popular machine learning library in Python.

Model Training and Evaluation

An experiment was logged using MLflow, allowing for easy tracking of model performance metrics and parameters.

Results

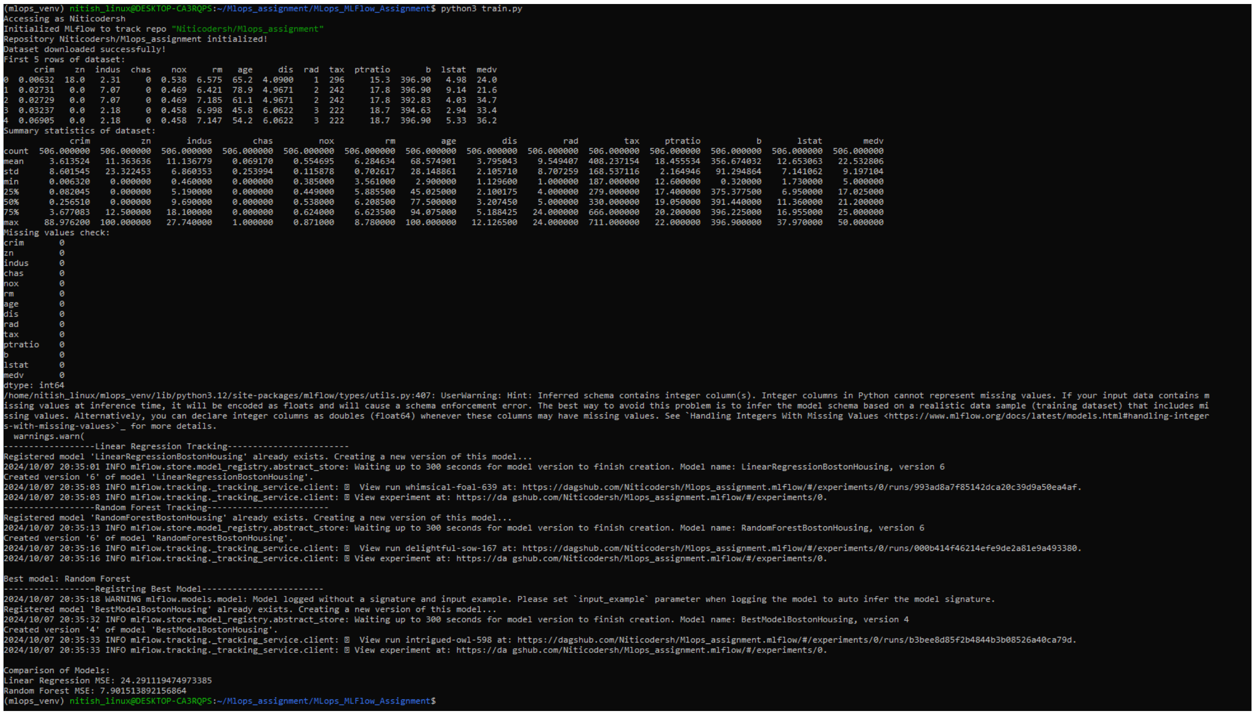
- **Linear Regression**:
  - Mean Squared Error (MSE): 24.29
  - R-squared: 0.67
- **Random Forest**:
  - Mean Squared Error (MSE): 7.90
  - R-squared: [Value]: 0.89

Conclusion

The performance of both models was evaluated using MSE and R-squared metrics. These metrics provide insight into the accuracy and reliability of the predictions made by each model. The use of MLflow facilitated the logging and tracking of experiments, ensuring reproducibility and ease of comparison between models.

ScreenShots

Locally running train.py, MLflow experiments logs are created on DagsHub (a remote server set as the tracking uri): https://dagshub.com/Niticodersh/Mlops_assignment.mlflow/#/experiments

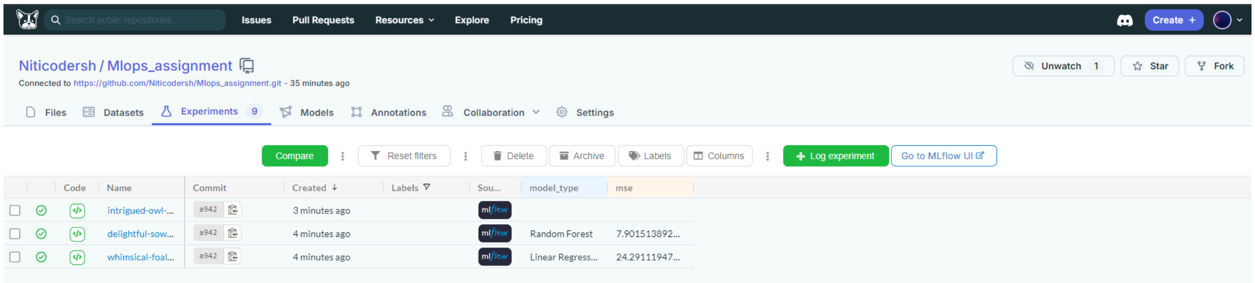


```
(mlops_venv) nitish_linux@DESKTOP-CA3RQPS:~/Mlops_assignment/MLops_MLFlow_Assignment$ python3 train.py
Accessing as Niticodersh
Initialized MLflow to track repo "Niticodersh/Mlops_assignment"
Repository Niticodersh/Mlops_assignment initialized!
Dataset downloaded successfully!
First 5 rows of dataset:
      crim    zn  indus  chas    nox     rm   age     dis  rad  tax  ptratio       b  lstat  medv
0  0.00632  18.0   2.31     0  0.538  6.575  65.2  4.0900    1  296     15.3  396.90   4.98  24.0
1  0.02731   0.0   7.07     0  0.469  6.421  78.9  4.9671    2  242     17.8  396.90   9.14  21.6
2  0.02729   0.0   7.07     0  0.469  7.185  61.1  4.9671    2  242     17.8  392.83   4.03  34.7
3  0.03237   0.0   2.18     0  0.458  6.998  45.8  6.0622    3  222     18.7  394.63   2.94  33.4
4  0.06905   0.0   2.18     0  0.458  7.147  54.2  6.0622    3  222     18.7  396.90   5.33  36.2
Summary statistics of dataset:
             crim          zn       indus        chas         nox          rm         age         dis         rad         tax     ptratio           b       lstat        medv
count  506.000000  506.000000  506.000000  506.000000  506.000000  506.000000  506.000000  506.000000  506.000000  506.000000  506.000000  506.000000  506.000000  506.000000
mean     3.613524   11.363636   11.136779    0.069170    0.554695    6.284634   68.574901    3.795043    9.549407  408.237154   18.455534  356.674032   12.653063   22.532806
std      8.601545   23.322453    6.860353    0.253994    0.115878    0.702617   28.148861    2.105710    8.707259  168.537116    2.164946   91.294864    7.141062    9.197104
min      0.006320    0.000000    0.460000    0.000000    0.385000    3.561000    2.900000    1.129600    1.000000  187.000000   12.600000    0.320000    1.730000    5.000000
25%      0.082045    0.000000    5.190000    0.000000    0.449000    5.885500   45.025000    2.100175    4.000000  279.000000   17.400000  375.377500    6.950000   17.025000
50%      0.256510    0.000000    9.690000    0.000000    0.538000    6.208500   77.500000    3.207450    5.000000  330.000000   19.050000  391.440000   11.360000   21.200000
75%      3.677083   12.500000   18.100000    0.000000    0.624000    6.623500   94.075000    5.188425   24.000000  666.000000   20.200000  396.225000   16.955000   25.000000
max     88.976200  100.000000   27.740000    1.000000    0.871000    8.780000  100.000000   12.126500   24.000000  711.000000   22.000000  396.900000   37.970000   50.000000
Missing values check:
crim       0
zn         0
indus      0
chas       0
nox        0
rm         0
age        0
dis        0
rad        0
tax        0
ptratio    0
b          0
lstat      0
medv       0
dtype: int64
/home/nitish_linux/mlops_venv/lib/python3.12/site-packages/mlflow/types/utils.py:407: UserWarning: Hint: Inferred schema contains integer column(s). Integer columns in Python cannot represent missing values. If your input data contains m
issing values at inference time, it will be encoded as floats and will cause a schema enforcement error. The best way to avoid this problem is to infer the model schema based on a realistic data sample (training dataset) that includes mi
ssing values. Alternatively, you can declare integer columns as doubles (float64) whenever these columns may have missing values. See `Handling Integers With Missing Values <https://www.mlflow.org/docs/latest/models.html#handling-integer
s-with-missing-values>`_ for more details.
  warnings.warn(
------------------Linear Regression Tracking------------------------
Registered model 'LinearRegressionBostonHousing' already exists. Creating a new version of this model...
2024/10/07 20:35:01 INFO mlflow.store.model_registry.abstract_store: Waiting up to 300 seconds for model version to finish creation. Model name: LinearRegressionBostonHousing, version 6
Created version '6' of model 'LinearRegressionBostonHousing'.
2024/10/07 20:35:03 INFO mlflow.tracking._tracking_service.client: ⍰  View run whimsical-foal-639 at: https://dagshub.com/Niticodersh/Mlops_assignment.mlflow/#/experiments/0/runs/993ad8a7f85142dca20c39d9a50ea4af.
2024/10/07 20:35:03 INFO mlflow.tracking._tracking_service.client: ⍰ View experiment at: https://da gshub.com/Niticodersh/Mlops_assignment.mlflow/#/experiments/0.
------------------Random Forest Tracking------------------------
Registered model 'RandomForestBostonHousing' already exists. Creating a new version of this model...
2024/10/07 20:35:13 INFO mlflow.store.model_registry.abstract_store: Waiting up to 300 seconds for model version to finish creation. Model name: RandomForestBostonHousing, version 6
Created version '6' of model 'RandomForestBostonHousing'.
2024/10/07 20:35:16 INFO mlflow.tracking._tracking_service.client: ⍰  View run delightful-sow-167 at: https://dagshub.com/Niticodersh/Mlops_assignment.mlflow/#/experiments/0/runs/000b414f46214efe9de2a81e9a493380.
2024/10/07 20:35:16 INFO mlflow.tracking._tracking_service.client: ⍰ View experiment at: https://da gshub.com/Niticodersh/Mlops_assignment.mlflow/#/experiments/0.

Best model: Random Forest
------------------Registring Best Model------------------------
2024/10/07 20:35:18 WARNING mlflow.models.model: Model logged without a signature and input example. Please set `input_example` parameter when logging the model to auto infer the model signature.
Registered model 'BestModelBostonHousing' already exists. Creating a new version of this model...
2024/10/07 20:35:32 INFO mlflow.store.model_registry.abstract_store: Waiting up to 300 seconds for model version to finish creation. Model name: BestModelBostonHousing, version 4
Created version '4' of model 'BestModelBostonHousing'.
2024/10/07 20:35:33 INFO mlflow.tracking._tracking_service.client: ⍰  View run intrigued-owl-598 at: https://dagshub.com/Niticodersh/Mlops_assignment.mlflow/#/experiments/0/runs/b3bee8d85f2b4844b3b08526a40ca79d.
2024/10/07 20:35:33 INFO mlflow.tracking._tracking_service.client: ⍰ View experiment at: https://da gshub.com/Niticodersh/Mlops_assignment.mlflow/#/experiments/0.

Comparison of Models:
Linear Regression MSE: 24.291119474973385
Random Forest MSE: 7.901513892156864
(mlops_venv) nitish_linux@DESKTOP-CA3RQPS:~/Mlops_assignment/MLops_MLFlow_Assignment$
```

Screenshots of Experiments performed hosted at DagsHub:

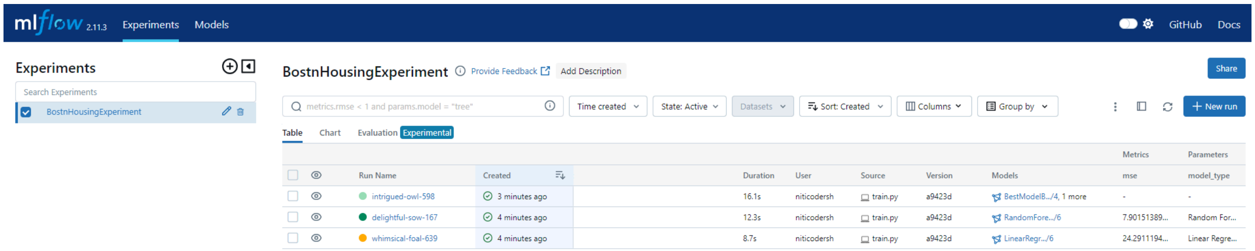


Niticodersh / Mlops_assignment

Connected to https://github.com/Niticodersh/Mlops_assignment.git - 35 minutes ago

		Code	Name	Commit	Created ↓	Labels	Sou...	model_type	mse
□	✓	</>	intrigued-owl-...	a942	3 minutes ago		mlflow		
□	✓	</>	delightful-sow...	a942	4 minutes ago		mlflow	Random Forest	7.901513892...
□	✓	</>	whimsical-foal...	a942	4 minutes ago		mlflow	Linear Regress...	24.29111947...

Screenshots of MLFlow UI hosted at DagsHub:

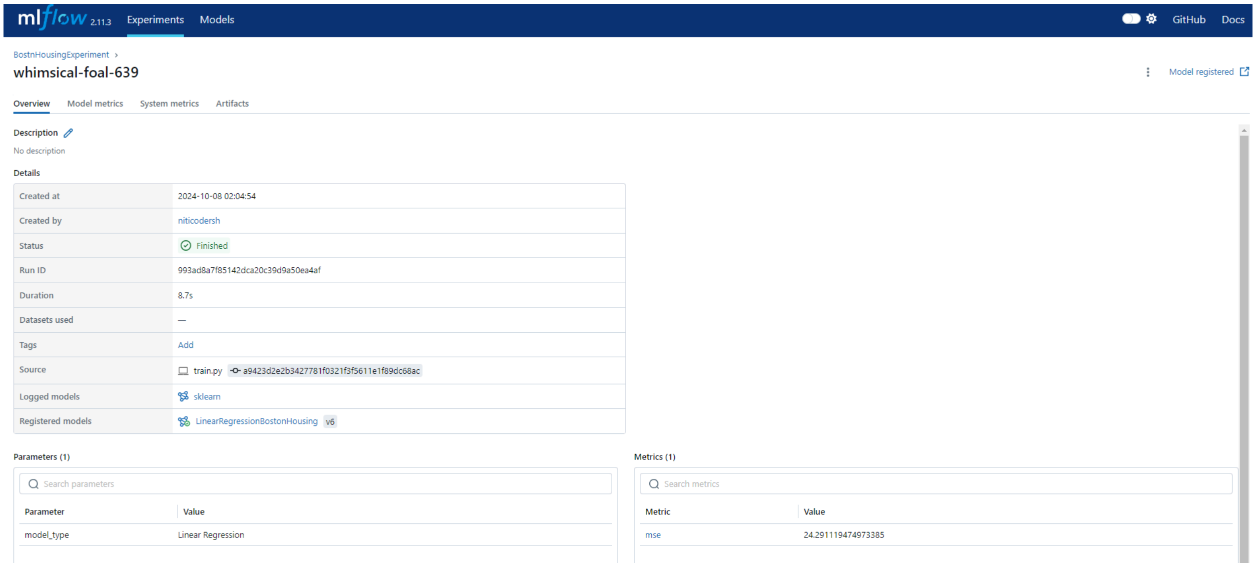


☑ BostnHousingExperiment

BostnHousingExperiment

									Metrics	Parameters
□		Run Name	Created	Duration	User	Source	Version	Models	mse	model_type
□		● intrigued-owl-598	✓ 3 minutes ago	16.1s	niticodersh	train.py	a9423d	BestModelB.../4, 1 more	-	-
□		● delightful-sow-167	✓ 4 minutes ago	12.3s	niticodersh	train.py	a9423d	RandomFore.../6	7.90151389...	Random For...
□		● whimsical-foal-639	✓ 4 minutes ago	8.7s	niticodersh	train.py	a9423d	LinearRegr.../6	24.2911194...	Linear Regre...

Linear Regression Model Results



BostnHousingExperiment ›

whimsical-foal-639

Description

No description

Details

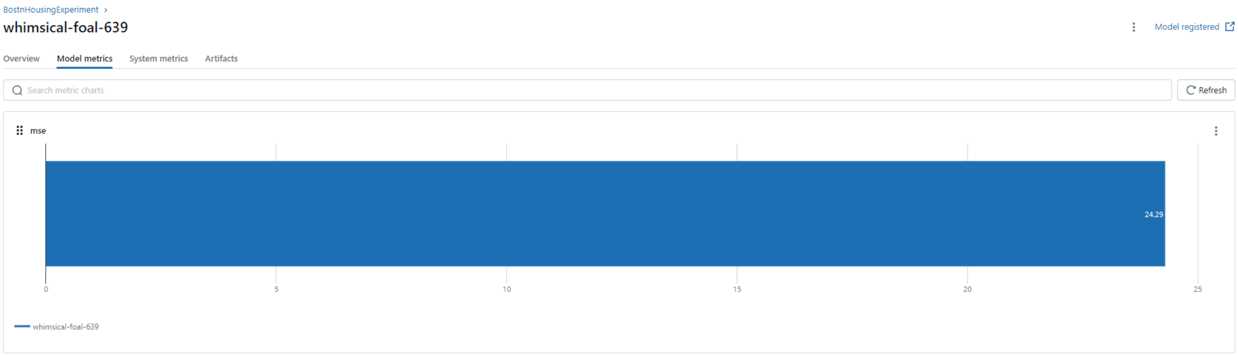
Created at	2024-10-08 02:04:54
Created by	niticodersh
Status	✓ Finished
Run ID	993ad8a7f85142dca20c39d9a50ea4af
Duration	8.7s
Datasets used	—
Tags	Add
Source	train.py a9423d2e2b3427781f0321f3f5611e1f89dc68ac
Logged models	sklearn
Registered models	LinearRegressionBostonHousing v6

Parameters (1)

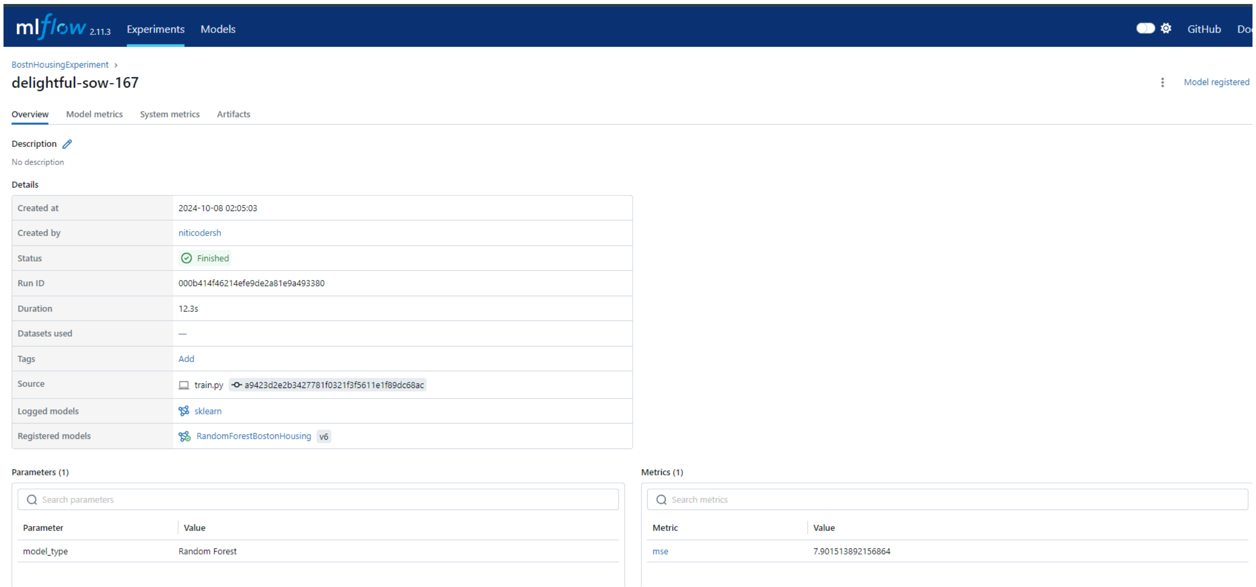
Parameter	Value
model_type	Linear Regression

Metrics (1)

Metric	Value
mse	24.291119474973385



Random Forest Model Results



delightful-sow-167

Description

No description

Details

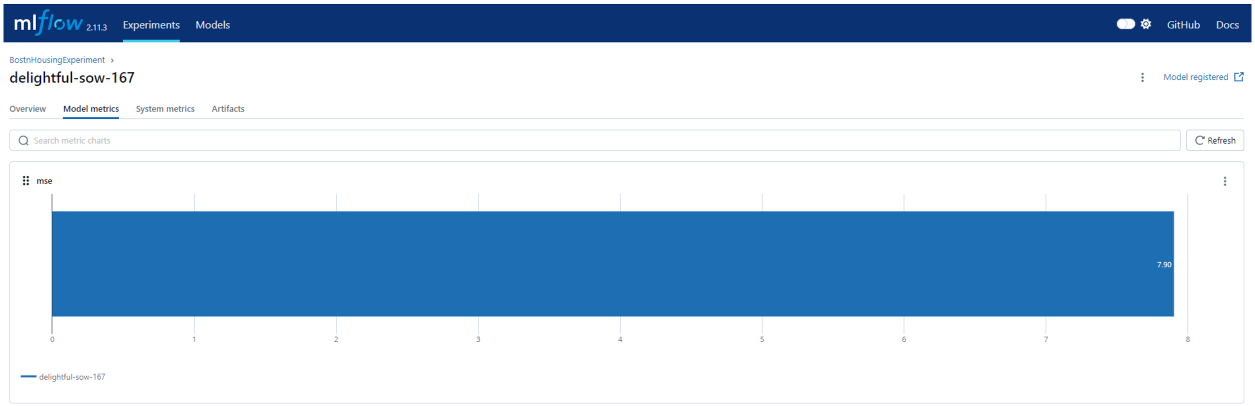
Created at	2024-10-08 02:05:03
Created by	niticodersh
Status	✓ Finished
Run ID	000b414f46214efe9de2a81e9a493380
Duration	12.3s
Datasets used	—
Tags	Add
Source	train.py a9423d2e2b3427781f0321f3f5611e1f89dc68ac
Logged models	sklearn
Registered models	RandomForestBostonHousing v6

Parameters (1)

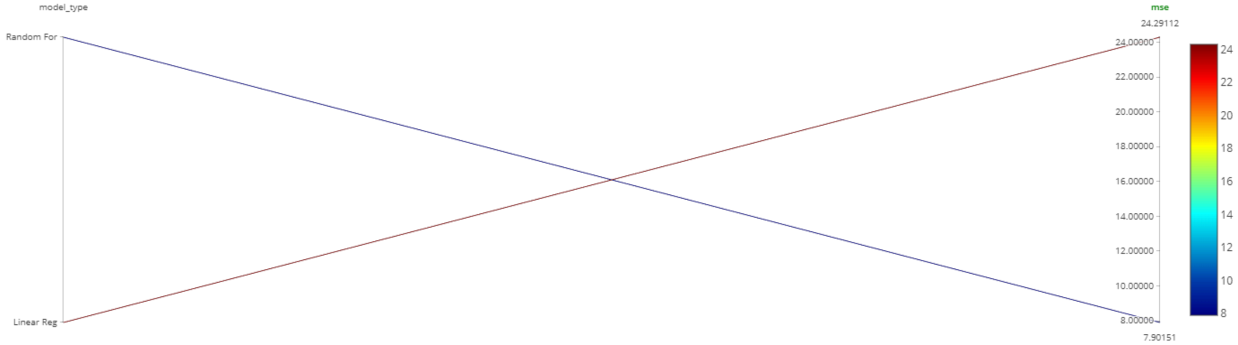
Parameter	Value
model_type	Random Forest

Metrics (1)

Metric	Value
mse	7.901513892156864



Both Models MSE Comparisons Result


model_type
Random For
Linear Reg
mse