REPORT

USA Housing – Linear Regression:

Introduction:

The dataset contains 5000 data points collected from a United States Housing.

Data Set Variables Information:

- **Area Income**
- **Area House Age**
- **Area No of Rooms**
- **Area No of Bedrooms**
- **Area Population**
- **Price**

In this dataset Area Income , Area House Age, Area No of Rooms , Area No of Bedrooms, and Area Population are X variable which dependent on Price and Price is the independent variable – y. Based on X variable we are predicting the Y Variable.

Objective of the Project:

- **Analyzing United States Housing – (US Housing) Dataset**
- **Analyzing models Linear Regression such as Multiple Linear Regression, Random Forest Regression Using Pyspark.**
- **Splitting the dataset into Training set and Test sets, Scaling the dataset**
- **Training and model on Test sets**
- **Performing R-Square, RMSE , P-Vales on Linear Regression model**
- **Performing Correlation along with Descriptive Analysis**
- **Performing K-Fold on Linear Regression models.**

Descriptive Analysis & Correlation

summary	count	mean	stddev	min	max
Area Income	5000	68583.10898397019	10657.991213888685	17796.63119	107701.7484
Area House Age	5000	5.977222035287008	0.9914561798324225	2.644304186	9.519088066
Area No of Rooms	5000	6.987791850909204	1.0058332312754115	3.236194023	10.75958834
Area No of Bedrooms	5000	3.9813299999999967	1.2341372654846832	2.0	6.5
Area Population	5000	36163.51603854035	9925.650113546026	172.6106863	69621.71338
Price	5000	1232072.6541452995	353117.6265836953	15938.65792	2469065.594

```
Correlation to Price for  Area Income 0.6397337782571293
Correlation to Price for  Area House Age 0.452542537178579
Correlation to Price for  Area No of Rooms 0.3356644533593983
Correlation to Price for  Area No of Bedrooms 0.1710710276560539
Correlation to Price for  Area Population 0.40855587932093074
Correlation to Price for  Price 1.0
```

Results of R-Square & RMSE Linear Regression model:

```
Note: the last rows are the information for Intercept
## -------------------------------------------------
##     Estimate   |   Std.Error | t Values  |  P-value
## -------------------------------------------------
##       21.508940     0.150804     142.628     0.000000
##   166230.977791  1618.070536     102.734     0.000000
##   120863.348440  1802.302245      67.061     0.000000
##     1393.558373  1470.612058       0.948     0.343388
##       15.057466     0.160901      93.582     0.000000
## -2631622.177511 19215.351741    -136.954     0.000000
## -------------------------------------------------
## Mean squared error: 10189018076.260710
## RMSE              : 100940.666118
## R-squared         : 0.917533
## Total iterations  : 1
```

Results of RMSE on Test Data after Cross Validation -Linear Regression model:

```
# cvModel uses the best model found from the Cross Validation
# Evaluate best model
print("Root Mean Squared Error (RMSE) on test data = %g" % rms)
```

```
Root Mean Squared Error (RMSE) on test data = 101768
```

R-Square of Random Forest model:

```
import sklearn.metrics
r2_score = sklearn.metrics.r2_score(y_true, y_pred)
print('r2_score: {:4.3f}'.format(r2_score))
```

```
r2_score: 0.920
```

RMSE of Random Forest model:

```
Root Mean Squared Error (RMSE) on test data = 228886
```

Best Parameters for Random Forest:

```
#BEST HYPERPARAMETERS

print('maxDepth - ', cvModel.bestModel._java_obj.getMaxDepth())
print('numTrees - ', cvModel.bestModel._java_obj.getNumTrees())
```

```
maxDepth -  6
numTrees -  20
```

Conclusion:

- R squared at 0.917 indicates that in our model, approximate 91.7% of the variability in "Price" can be explained using the model. It is a very good model.
- RMSE measures the differences between predicted values by the model and the actual values. After analyzing with the actual "**Price**" value, with such as mean, min and max. After such comparison, our RMSE looks Very good.
- If we analyze the R-Square values of two models, we notice that the Random Forest has the highest values of 92% than followed by Multiple Linear Regression with 91.7%. Therefore, we can conclude both the models are very good.