

HOUSING: PRICE PREDICTION

Submitted by:

NAMRATA BARASKAR

**ACKNOWLEDGMENT**

The internship opportunity I had with Flip Robo was a great chance for learning and professional development. Therefore, I consider myself as a very lucky individual as I was provided with an opportunity to be a part of it.

I would like to thank our SME for suggesting this project and for his whole hearted cooperation and constant encouragement throughout the project.

And I also like to thank the data trained mentors and Technical team members for helping me with the technical queries.

And these are the following website which I referred for the reference

1. <https://www.kaggle.com/>
2. <https://scikit-learn.org/>
3. [www.stackoverflow.com](http://www.stackoverflow.com)
4. [www.google.com](http://www.google.com)
5. [www.geeksforgeeks.org](http://www.geeksforgeeks.org)

**INTRODUCTION**

* Business Problem Framing
* Houses are one of the necessary need of each and every person around the globe and therefore housing and real estate market is one of the markets which is one of the major contributors in the world’s economy. It is a very large market and there are various companies working in the domain. Data science comes as a very important tool to solve problems in the domain to help the companies increase their overall revenue, profits, improving their marketing strategies and focusing on changing trends in house sales and purchases. Predictive modelling, Market mix modelling, recommendation systems are some of the machine learning techniques used for achieving the business goals for housing companies. Our problem is related to one such housing company. A US-based housing company named Surprise Housing has decided to enter the Australian market. The company uses Logistic Regression is a part of the Supervised Learning method of Machine Learning. It is a statistical method for the analysis of a dataset. It has one or more independent variables that determine an outcome. There is one basic difference between Linear Regression and Logistic Regression  which is that Linear Regression's outcome is continuous whereas Logistic Regression's outcome is only limited. Here, the outcome represents a dependent variable.
* Conceptual Background of the Domain Problem
* In real estate the value of property usually increases with time as seen in many countries. One of the causes for this is due to rising population.
* The value of property also depends on the proximity of the property, its size its neighbourhood and audience for which the property is subjected to be sold. For example if audience is mainly concerned of commercial purpose. Then the property which is located in densely populated area will be sold very fast and at high prices compared to the one located at remote place. Similarly if audience is concerned only on living place then property with less dense area having large area with all services will be sold at higher prices.
* The company is looking at prospective properties to buy houses to enter the market. We are required to build a model using Machine Learning in order to predict the actual value of the prospective properties and decide whether to invest in them or not.
* Review of Literature

From the dataset description I have leant the nature of data.

* Motivation for the Problem Undertaken
* To understand real world problems where Machine Learning and Data Analysis can be applied to help organizations in various domains to make better decisions with the help of which they can gain profit or can be escaped from any loss which otherwise could be possible without the study of data .One of such domain is Real Estate.
* Houses are one of the necessary need of each and every person around the globe and therefore housing and real estate market is one of the markets which is one of the major contributors in the world’s economy. It is a very large market and there are various companies working in the domain. Data science comes as a very important tool to solve problems in the domain to help the companies increase their overall revenue, profits, improving their marketing strategies and focusing on changing trends in house sales and purchases. Predictive modelling, Market mix modelling, recommendation systems are some of the machine learning techniques used for achieving the business goals for housing companies. Our problem is related to one such housing company.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

[Regardless of where you stand on the matter of Data Science sexiness, it’s simply](https://www.glassdoor.com/Best-Jobs-in-America-LST_KQ0%2C20.htm)

[impossible to ignore the continuing importance of data, and our ability to analyze,](https://www.glassdoor.com/Best-Jobs-in-America-LST_KQ0%2C20.htm) [organize, and contextualize it. Drawing on their vast stores of employment data and](https://www.glassdoor.com/Best-Jobs-in-America-LST_KQ0%2C20.htm) [employee feedback, Glassdoor ranked Data Scientist #1 in their 25 Best Jobs in](https://www.glassdoor.com/Best-Jobs-in-America-LST_KQ0%2C20.htm) [America list. So the role is here to stay, but unquestionably, the specifics of what a](https://www.glassdoor.com/Best-Jobs-in-America-LST_KQ0%2C20.htm) [Data Scientist does will evolve. With technologies like Machine Learning becoming](https://www.glassdoor.com/Best-Jobs-in-America-LST_KQ0%2C20.htm) [ever-more common place, and emerging fields like Deep Learning gaining significant](https://www.glassdoor.com/Best-Jobs-in-America-LST_KQ0%2C20.htm) [-](https://www.glassdoor.com/Best-Jobs-in-America-LST_KQ0%2C20.htm) [Data Scientists continue to ride the crest of an incredible wave of innovation and](https://www.glassdoor.com/Best-Jobs-in-America-LST_KQ0%2C20.htm) [technological progress.](https://www.glassdoor.com/Best-Jobs-in-America-LST_KQ0%2C20.htm)

While having a strong coding ability is important, data science isn’t all about

software engineering (in fact, have a good familiarity with Python and you’re good to go). Data scientists live at the intersection of coding, statistics, and critical thinking. As Josh Wills put it, “data scientist is a person who is better at statistics than any programmer and better at programming than any statistician.” I personally know too many software engineers looking to transition into data scientist and blindly utilizing machine learning frameworks such as TensorFlow or Apache Spark to their data without a thorough understanding of statistical theories behind them. So comes the study of statistical learning, a theoretical framework for machine learning drawing from the fields of statistics and functional analysis.

Why study Statistical Learning? It is important to understand the ideas behind the

various techniques, in order to know how and when to use them. One has to understand the simpler methods first, in order to grasp the more sophisticated ones. It is important to accurately assess the performance of a method, to know how well or how badly it is working. Additionally, this is an exciting research area, having important applications in science, industry, and finance. Ultimately, statistical learning is a fundamental ingredient in the training of a modern data scientist

* Data Sources and their formats

Data I get form the Flip Robo the format was in CSV (Comma Separated Values).The number of columns and row are 1168 and columns are 81.

* Data Pre-processing Done

The raw data is taken and performed various steps to reduce skewness, outlier, class imbalance and scaling. There were no null value was present in the dataset but there are some outliers which also get too removed. Many outlier removal and skewness removal methods are tested and best method Is chosen in order to prevent data loss.

* Hardware and Software Requirements and Tools Used

Hardware – Laptop

Software – google colab, jupyter notebook

Libraries- numpy, pandas, seaborn, matplotlib.pyplot,sklearn.

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

Classification Model with following algorithms

* Linear Regression
* Random forest Regression
* DecisionTree Regression

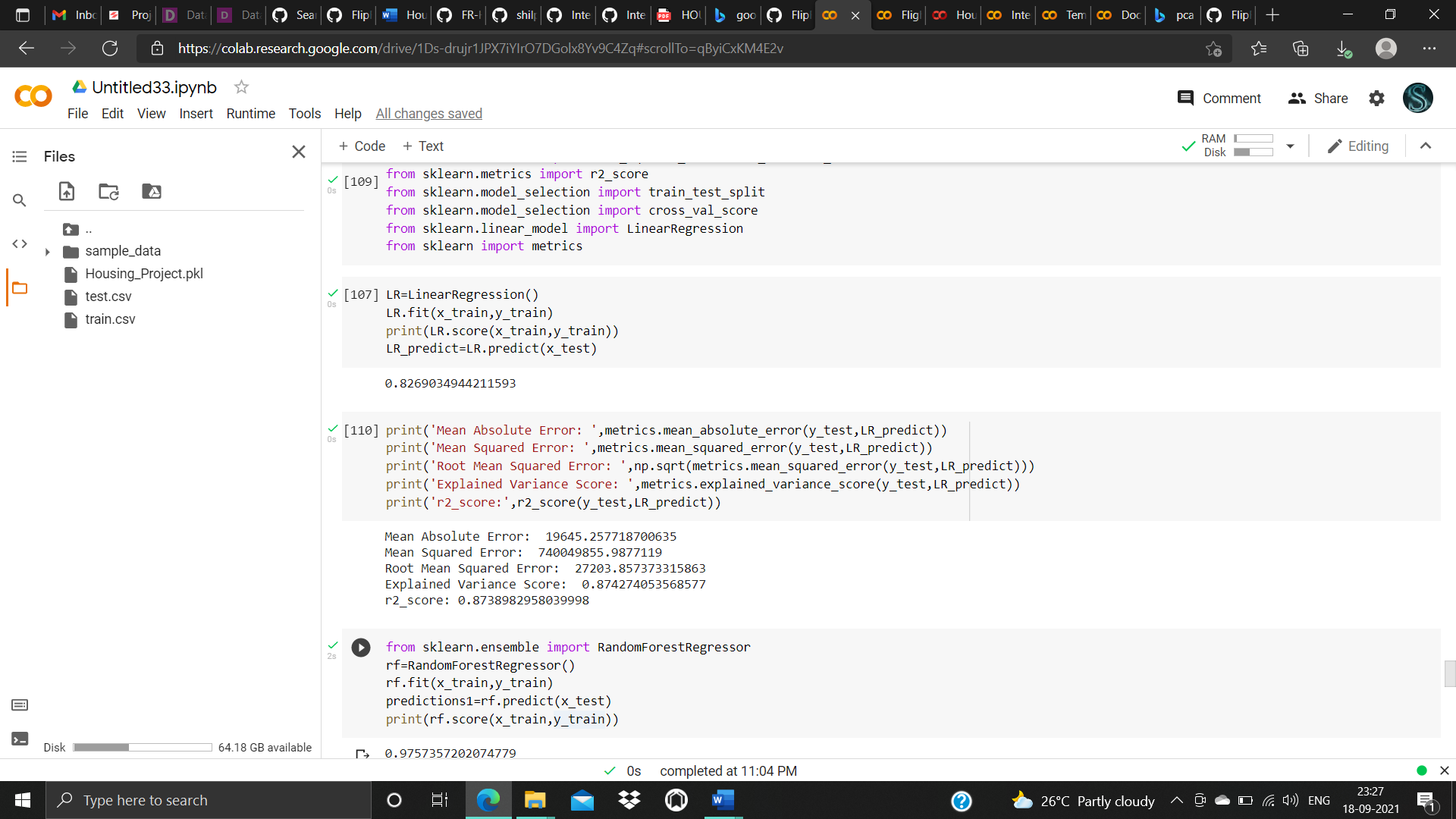
Evaluation metrics

* Mean Absolute error
* Mean square error
* Root mean squared error
* Varance
* R2 score
* Testing of Identified Approaches (Algorithms)

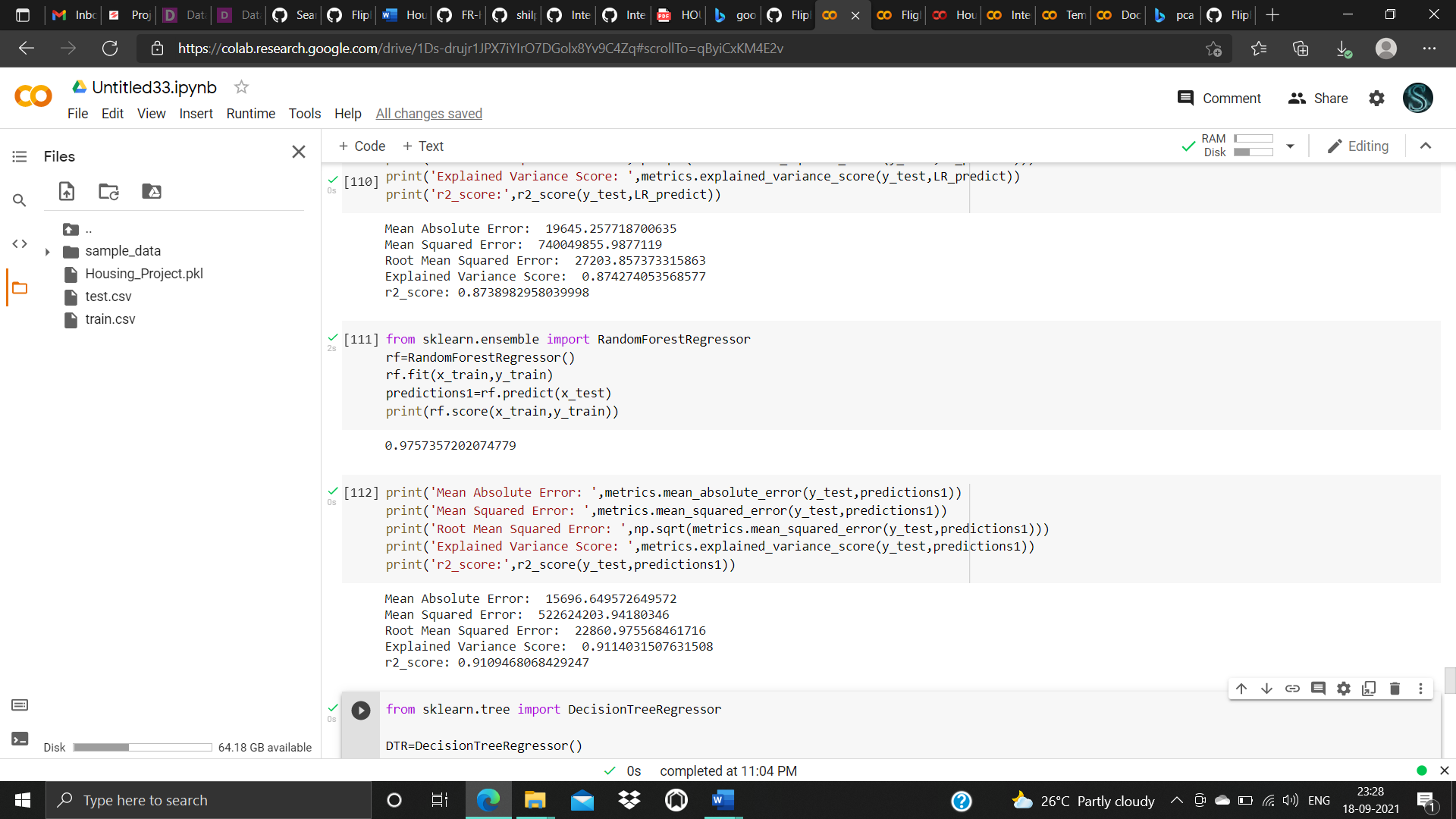
Listing down all the algorithms used for the training and testing.

* LR=LinearRegression()
* DT=DecisionTreeRegressor()
* RF=RandomForestRegressor()
* Run and Evaluate selected models

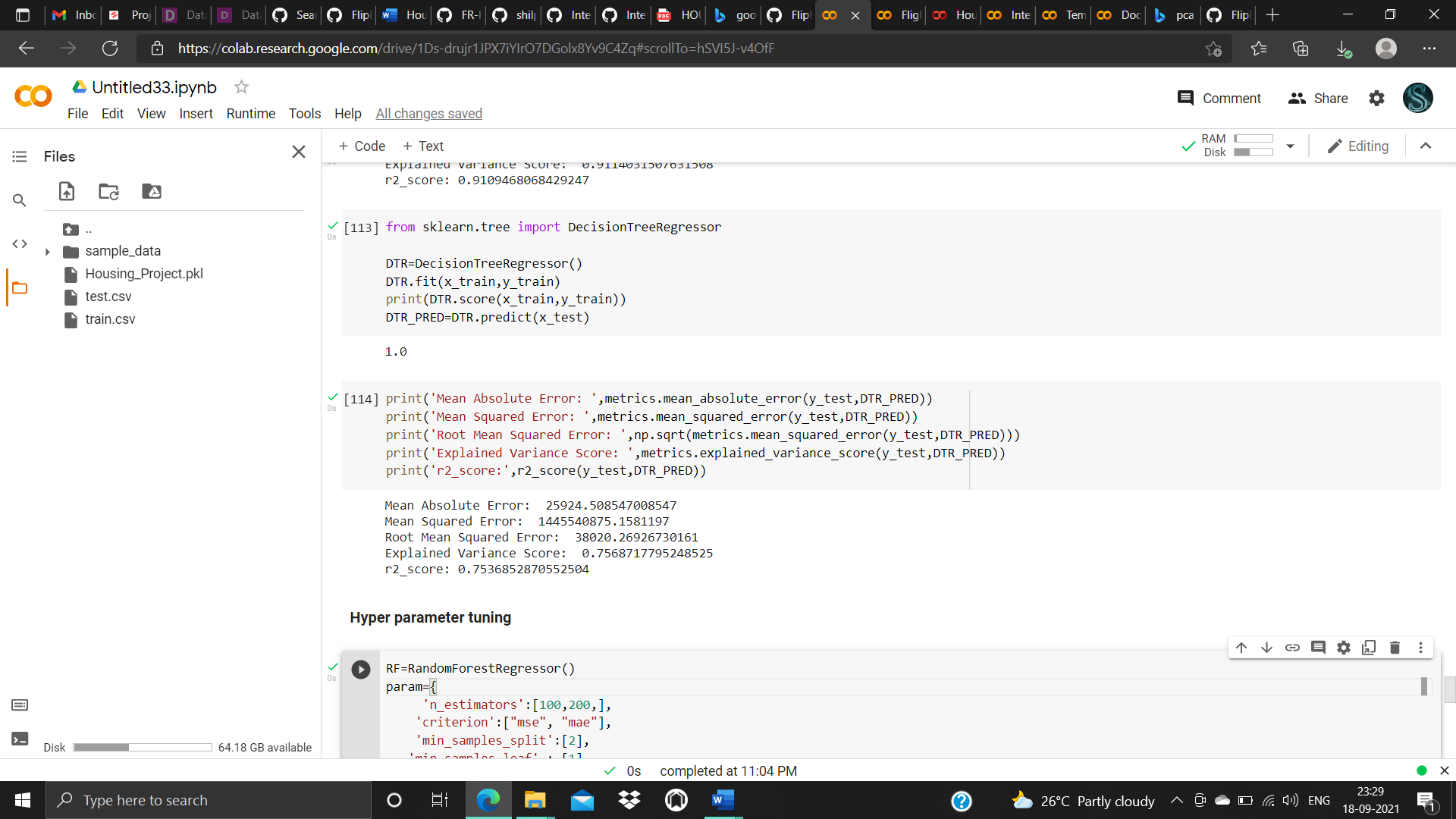
Describe all the algorithms used along with the snapshot of their code and what were the results observed over different evaluation metrics.



1.Linear Regressor

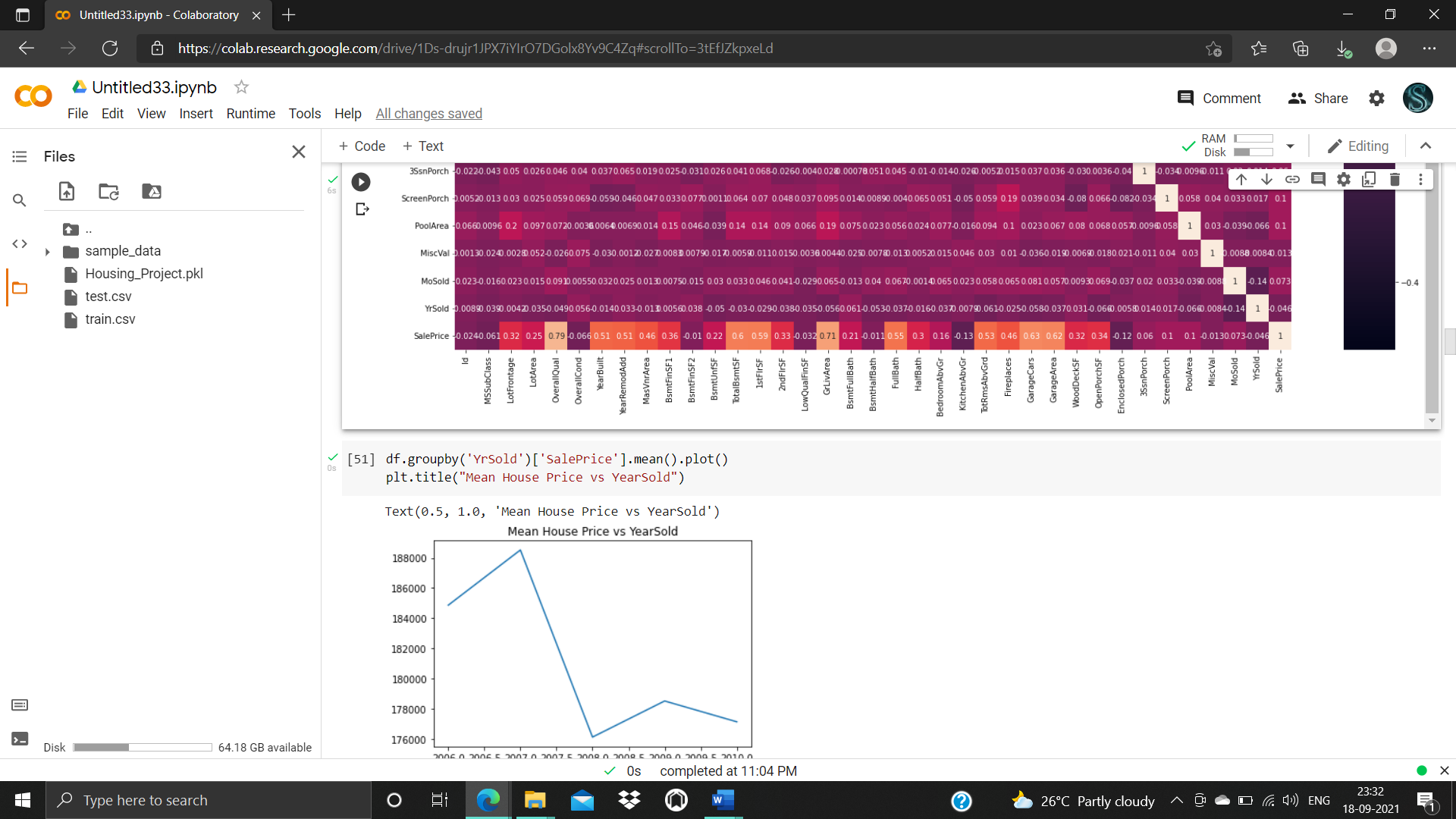


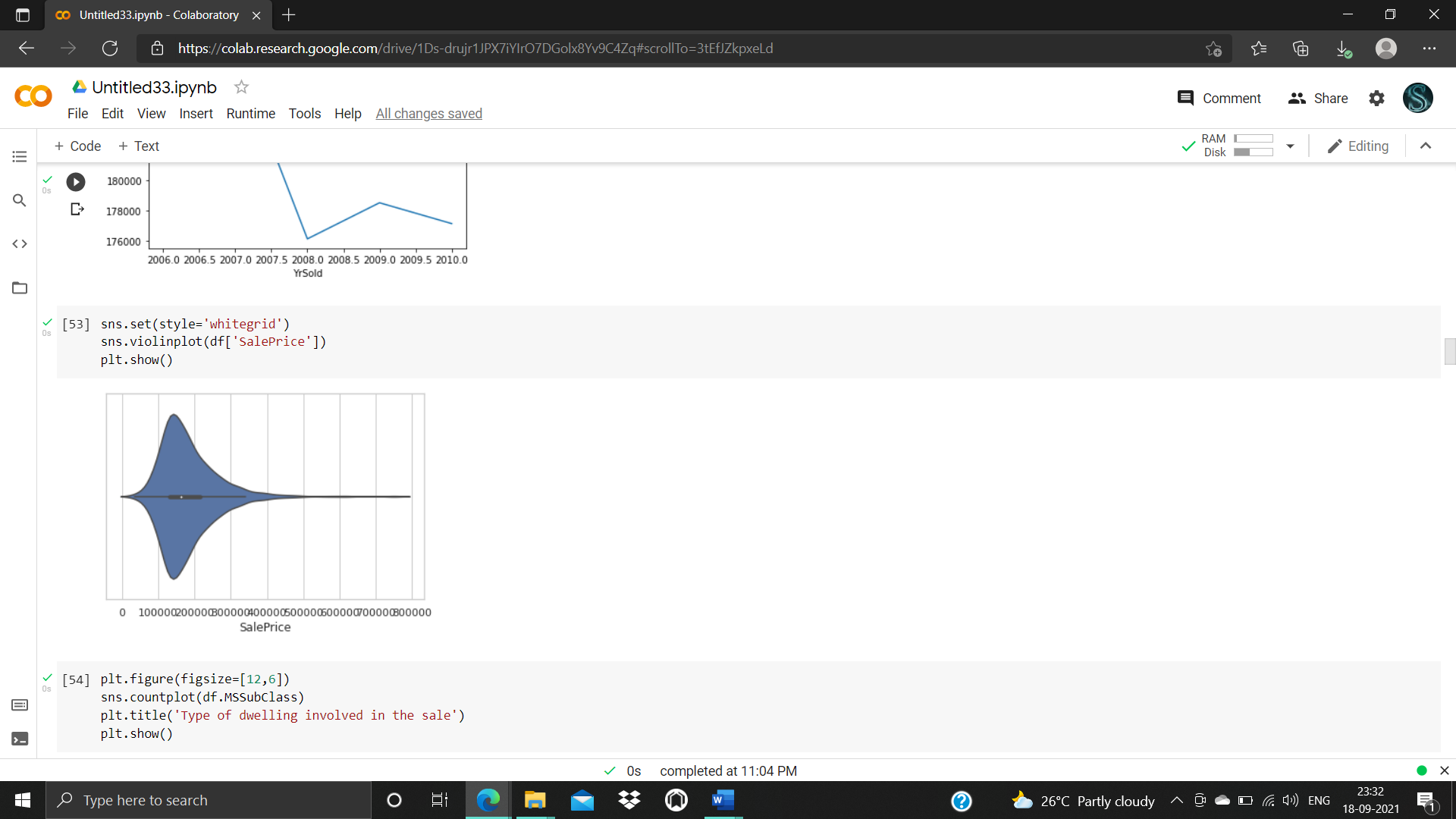
2.Random Forest Regressor

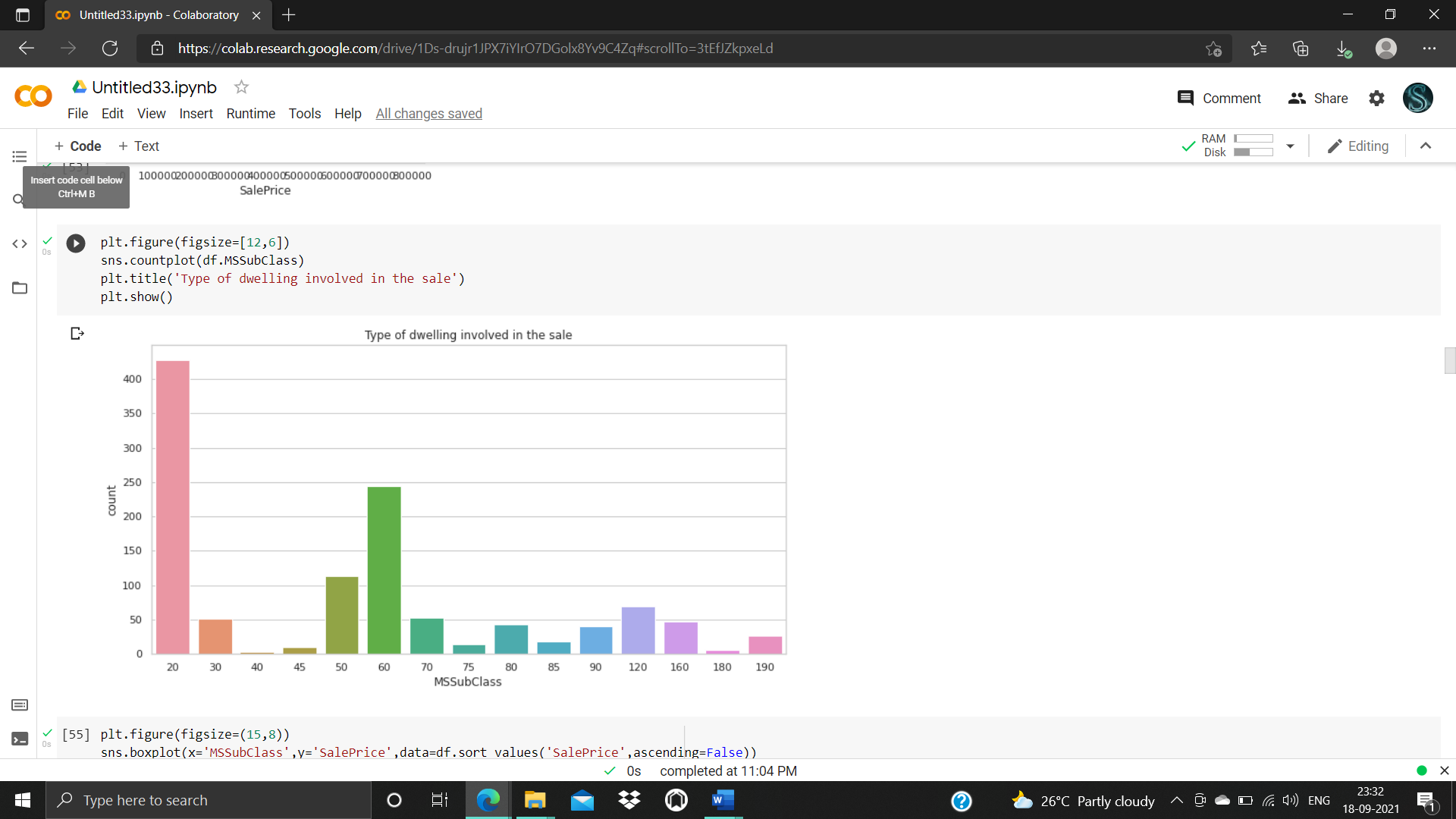


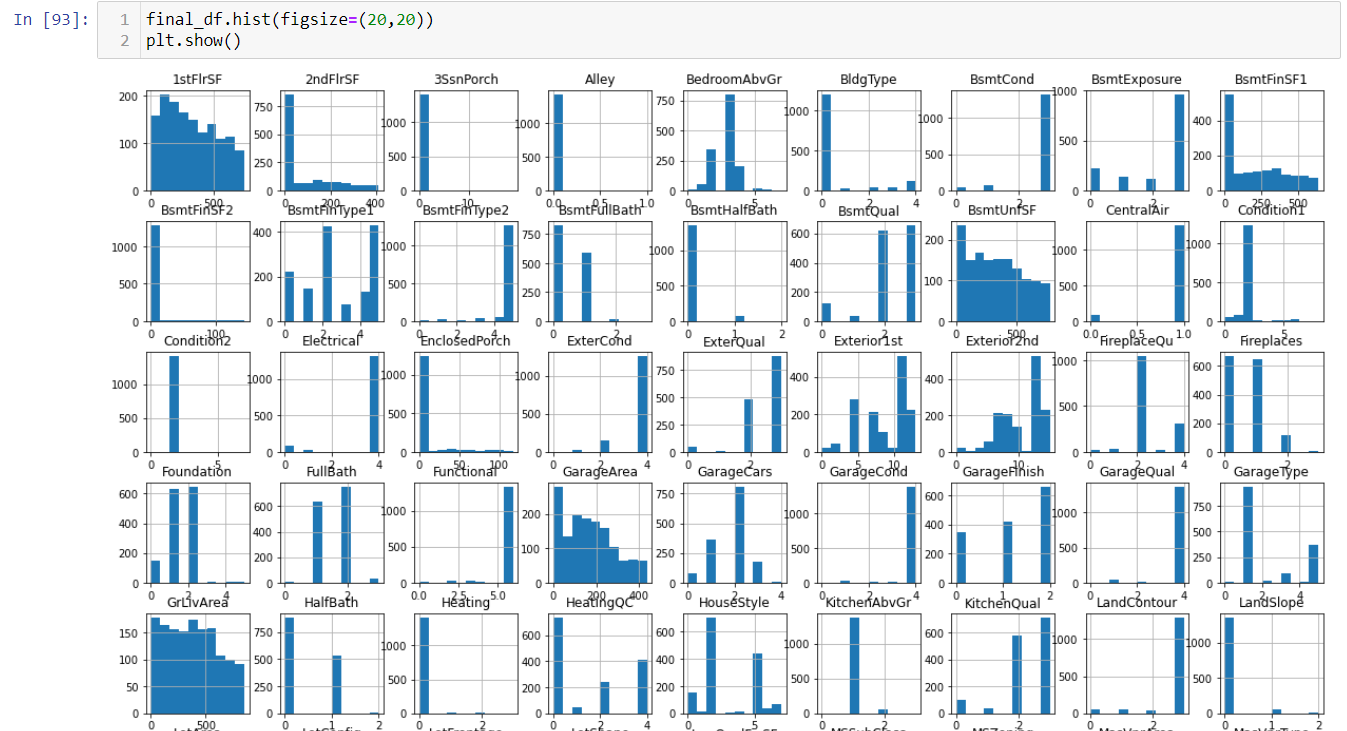
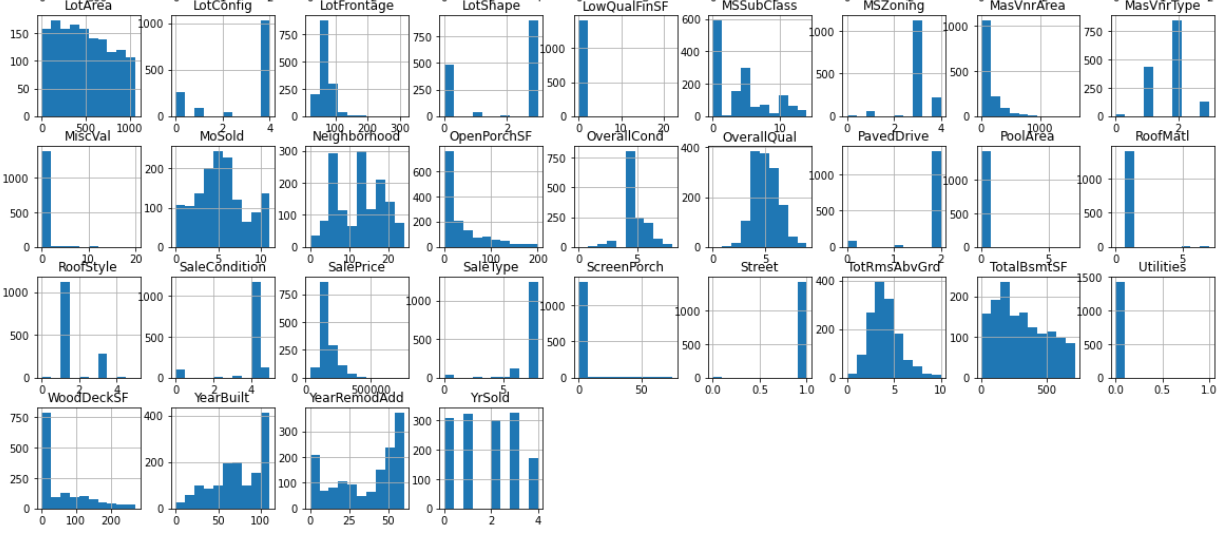
3. Decision Tree

* Key Metrics for success in solving problem under consideration
* Mean Absolute error
* Mean square error
* Root mean squared error
* Varance
* R2 score
* Visualizations







* 
* 
* Interpretation of the Results

Data is highly skewed from the mean so skewness removal methods need to be followed.

Box plot tells that there are many outliers are present in the data so need to remove the outliers too.

Heat map tells that which data is highly correlated with class are more important in constructing the model.

**CONCLUSION**

Logistic Regression is a very good part of Machine Learning. It is used in various fields, like medical, banking, social science, etc. It can predict the value based on the training dataset. The training dataset defines it accurately.

* RandomForest: A Random Forest is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees and a technique called Bootstrap and Aggregation, commonly known as bagging. The basic idea behind this is to combine multiple decision trees in determining the final output rather than relying on individual decision trees.Random Forest has multiple decision trees as base learning models. We randomly perform row sampling and feature sampling from the dataset forming sample datasets for every model. This part is called Bootstrap.We need to approach the Random Forest regression technique like any other machine learning technique
* Design a specific question or data and get the source to determine the required data.
* Make sure the data is in an accessible format else convert it to the required format.
* Specify all noticeable anomalies and missing data points that may be required to achieve the required data.
* Create a machine learning model
* Set the baseline model that you want to achieve
* Train the data machine learning model.
* Provide an insight into the model with test data
* Now compare the performance metrics of both the test data and the predicted data from the model.
* If it doesn’t satisfy your expectations, you can try improving your model accordingly or dating your data or use another data modeling technique.
* At this stage you interpret the data you have gained and report accordingly.
* Key Findings and Conclusions of the Study

From this dataset I get to know that each feature play a very import role to understand the data. Data format plays a very important role in the visualization and Appling the models and algorithms. Importance of removing the skewness and outlier is important. Finding the best parameters for the algorithm also plays a important role in performance and accuracy of the model.

* Learning Outcomes of the Study in respect of Data Science

Learnt how to process the large number of data. Tried and learnt more about distribution of the data. The power of visualization is helpful for the understanding of data into the graphical representation its help me to understand that what data is trying to say, Data cleaning is one of the most important step to remove missing value or null value fill it by mean median or by mode or by 0.Setting a good parameters is more important for the model accuracy. Finding a best random state played a vital roll in finding a better model.

* Limitations of this work and Scope for Future Work

The techniques to increase the speed of the model need to be constructed. The future model can be constructed with the most co related data with the target variable in order to increase the speed of the model.