Cognizant

Welcome
to the course

Linear Regression

Linear Regression

Is the Property Worth It?

Meet Leslie Wong, the owner of the real estate firm, Wong's World. He is on his way to meet a client to buy a large property in downtown Singapore.

A lot depends on this deal for Wong and his company. This is the first time that his company is investing a huge sum of money.

If they can get this deal right, they will certainly be one of the major real estate companies in the city. But they have a lot of task at hand now; and they have to do it fast before other agents flock in.

Linear Regression

Is the Property Worth It?

These are daunting tasks. To get these done, Leslie has hired Liam, a data scientist by profession. His task is to help Wong's World take a profitable decision.

Linear Regression

Is the Property Worth It?

As they travel together, Leslie could not stop but ask Liam about his strategy to address the problem.

Linear Regression

Objectives

By the end of this module, you will be able to:

- Define Linear Regression.
- Identify the methods of selecting data for Linear Regression.
- Explain data analysis features such as plotting, outliers, and seasonality.
- Define machine learning models.
- Explain how to analyze dataset for the machine learning model.
- Recognize the error handling methods.
- Explain the process to apply the dataset into the machine learning model.

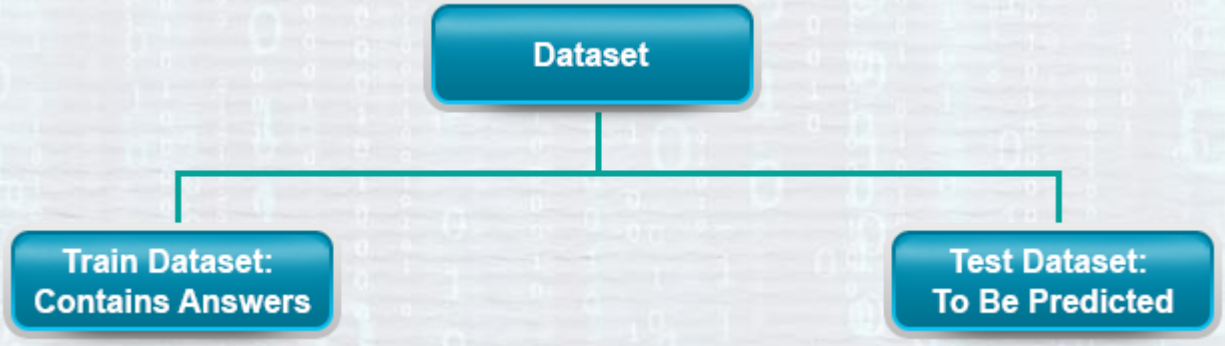
Linear Regression

Linear Regression

Liam's task is to find out the best resale price of the house. He can execute this task using the Linear Regression model. The basic premise of a Linear Regression model is Linear Algebra. Linear Regression is used for predicting continuous valued output and its outcome can have any one of an infinite number of possible values.

Linear Regression is a Supervised Learning algorithm. The input to a Supervised Learning model is a labeled dataset that has the right answers given. A linear regression model that involves a single variable is known as Univariate linear regression. In the current context, details of houses such as location, the sq. meter of the houses, and their resale price are the input data to the Algorithm or the training dataset.

The test dataset contains the details of houses for which the best possible resale price needs to be predicted.



Cognizant

Linear Regression

Dataset

Liam has researched a lot on the data to be used as a labeled dataset. A couple of days back, he eventually found a long list of properties in Singapore, along with their resale prices. The list contains 6581 rows. The details are provided under various columns.

Let's view the details of the columns in the dataset Liam found.

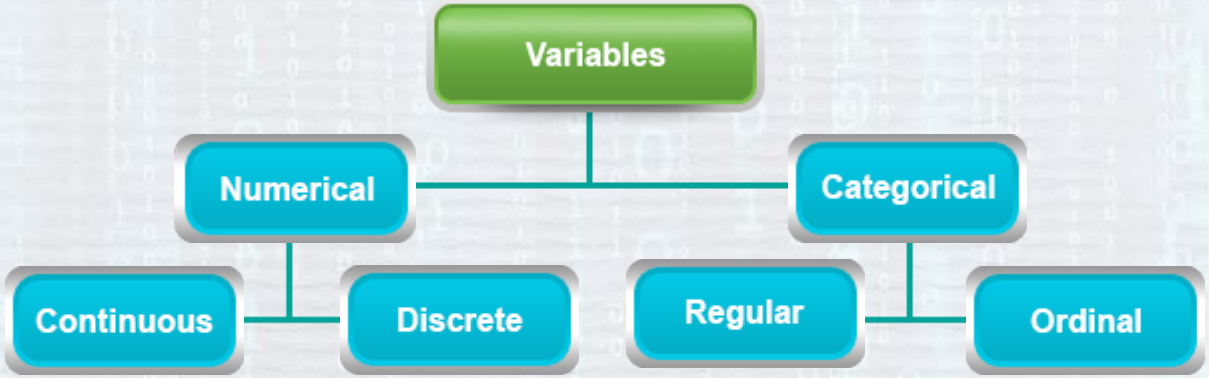
Note: The resale price value is the response variable in this case.

Columns	Description
Area	Location of the house
Block	Block number
Address	Street address
Floor	Floor number in the block
SQMS	Size of the flat in square meters
LeaseCommenceDate	Date on which the house was leased out
Approval Date	Date on which the resale was approved
FlatType	Type of flat (e.g. improved, modeled)
Bedrooms	No. of bedrooms in the house
ResalePrice	Resale price of the house

Linear Regression

Understanding Variables

Liam's next task is to identify the variables. He refers to the chart depicting the list of variables and tries to identify what types of variables are available in the dataset.



Cognizant

Linear Regression

Understanding Variables

Numerical Variable

It can take a wide range of numerical values, and it is sensible to add, subtract, or take averages with those values.

Categorical Variable

It consists of categories. Possible values are called the variable's levels.

Continuous Variable

It takes any numerical value including negatives and decimals. Example: ResalePrice.

Discrete Variable

It takes only whole, non-negative numbers such as 0,1, and 2. Example: Bedrooms.

Regular Variable

It includes all categorical variables that are non-ordinal. Example: FlatType.

Ordinal Variable

It has categorical variables, but its levels have a natural ordering. Example: Low, medium, and high.

Linear Regression

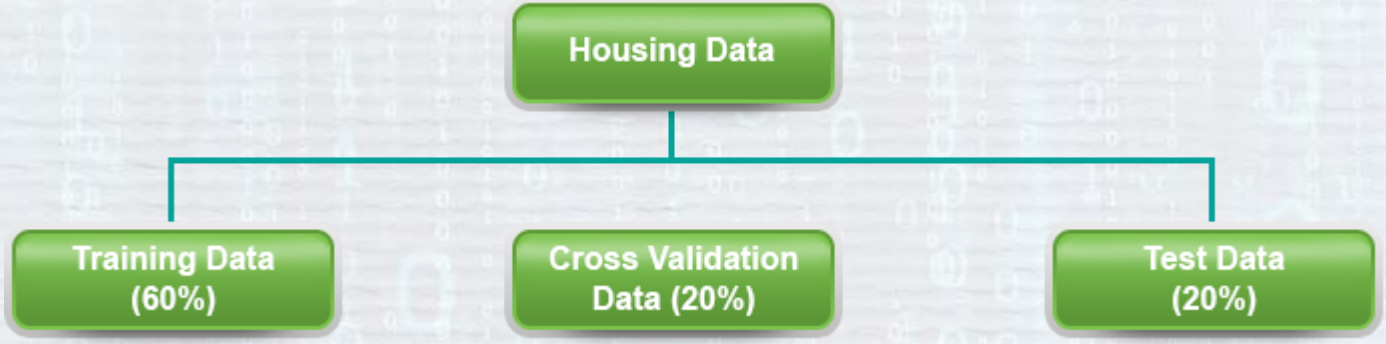
Data Selection Strategy for Training

As Liam explains the dataset, Leslie seems to be lost in the immense data. Liam assures him that he would take only the relevant data and modify it for the learning model. He would use the data selection strategy for training.

To do that, Liam needs to create three different samples of the training data having the following proportions.

- 60% for training the model
- 20% for cross validating the model
- Last 20% for finalizing the model

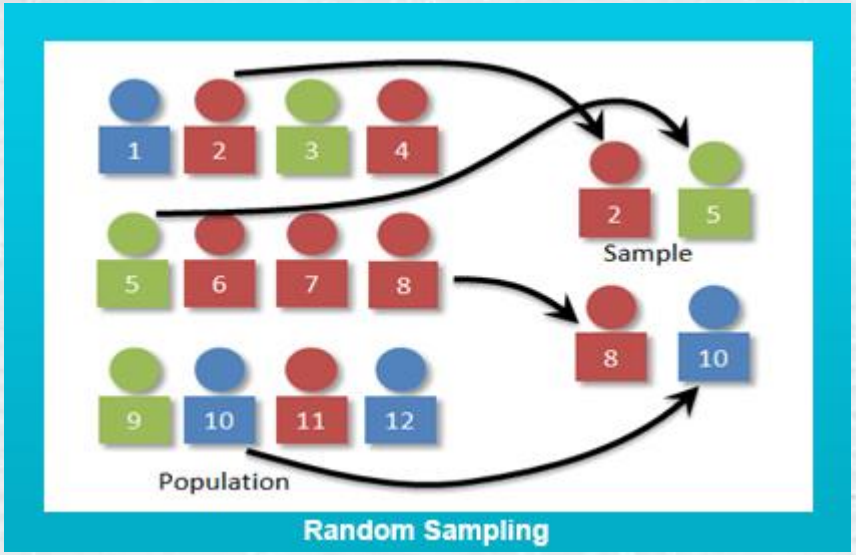
Refer to the diagram to view a visual representation of the dataset.

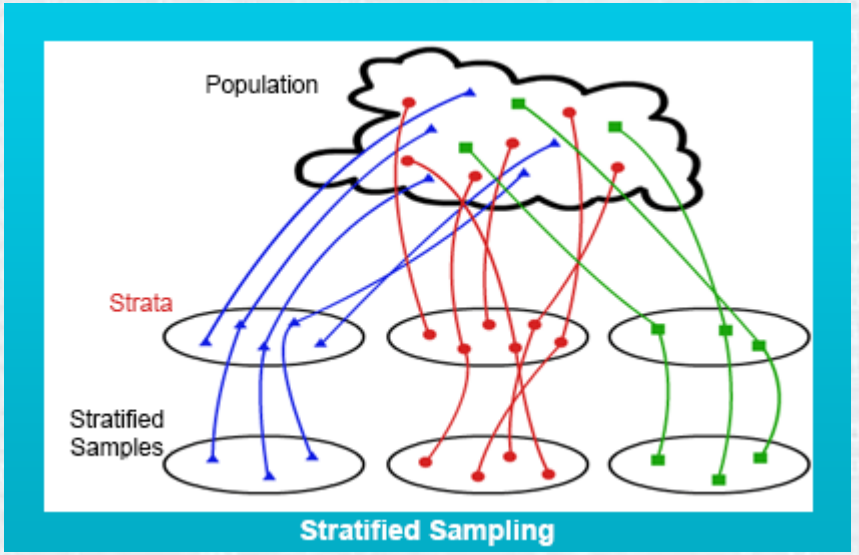


Linear Regression

Data Sampling

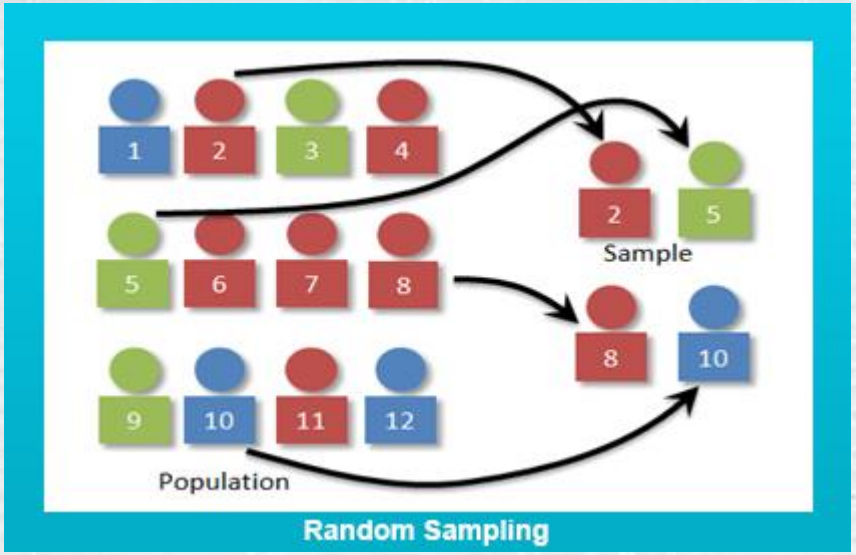
Liam can distribute the data into the three buckets of data sampling in one of the two ways: Random Sampling and Stratified Sampling.





Linear Regression

Data Sampling



Random Sampling

In random sampling, a sample is selected so that each item or person in the population has the same chance of being included. Thus, it refers to the process rather than the outcome of the process. This method assembles samples easily.

For example, if you want to draw a sample of five items for a group of 50 using random sampling, you can place them in a hat and draw five of them randomly.

However, random sampling may not present the proportion present in the original database. So, if the population of a city can be separated into smaller groups based on one or more distinguishing characteristics, random sampling may not be the best solution.

Linear Regression

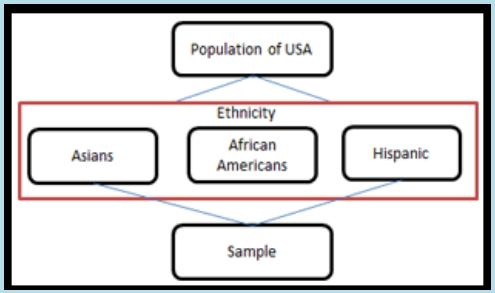
Data Sampling

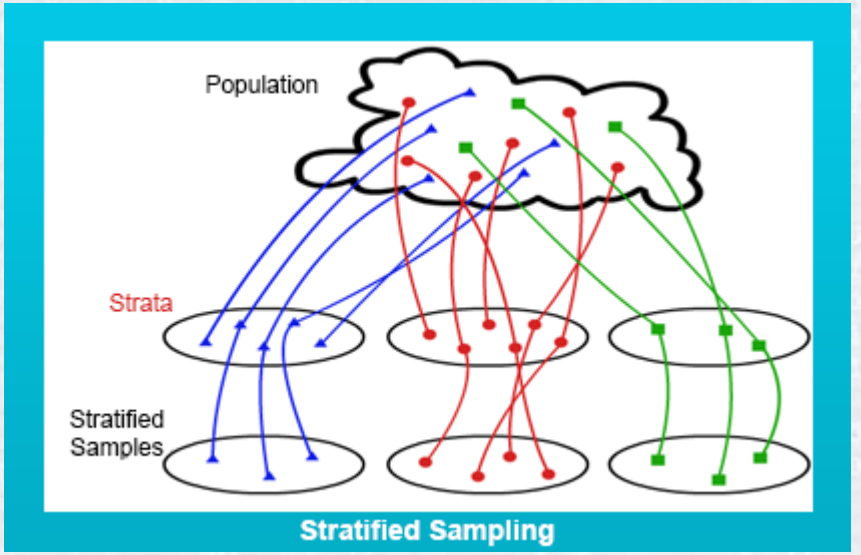
Stratified Sampling

In stratified sampling, elements in the dataset are first divided into various strata. Random sample is then taken from each stratum. A strata column is a categorical column with discrete values. For example, you can divide the population of USA into various ethnic groups. These groups will serve as strata.

Stratification ensures that the ratios of the selected values are preserved. Therefore, stratified techniques are generally used when population can be separated into smaller groups based on one or more distinguishing characteristics.

In Leslie's case, stratified sampling can be very effective while selecting some of the data.





Linear Regression

Data Selection Strategy for Training: Example

In the current context, the Area column is an important feature and analysis indicates the variation in the Resale Price is based on Area.

Area can be a potential strata column so that the original proportion of the sample is maintained across the sub-samples of training, cross validation, and test dataset.

Take a look at the diagram to see an example of how the dataset can be arranged into strata columns.

Linear Regression

Check Your Understanding

You have come across a lot of data on the type of houses in your city (e.g. flats, bungalows, houses, commercial buildings etc.). Using this data you have to prepare a chart showing the popularity of various building types in your city.

What sampling strategy should be the most effective for this task?

- Random Sampling
- Stratified Sampling

Linear Regression

Check Your Understanding

Correct Answer:

Stratified Sampling

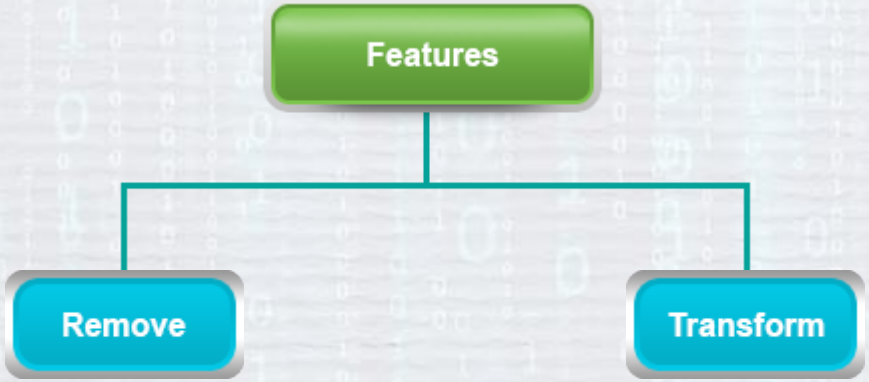
Point to Remember

Stratified sampling is the most effective sampling strategy for this task. First you need to create strata columns based on the building types to ensure that they are represented in the right proportion.

Linear Regression

Understanding Data

Liam now starts the data analysis. His primary concern at this point is to find out whether a feature should be removed, transformed, or taken forward.



Linear Regression

Understanding Data

Remove

The following features can be removed as they don't hold any key to the resale price of the house:

- Block Number (alphanumeric data point)
- Address (combination of text and number)

Transform

The following features can be transformed into new features.

- LeaseCommenceDate: It can be transformed into the feature Age by subtracting a constant value 2015. The output would be the age of the apartment in number of years.
- Approval Date: It can be transformed into a new feature, DaysElapsedSinceTransaction by subtracting the current date from the Approval Date. The output would be the days elapsed since the resale happened.

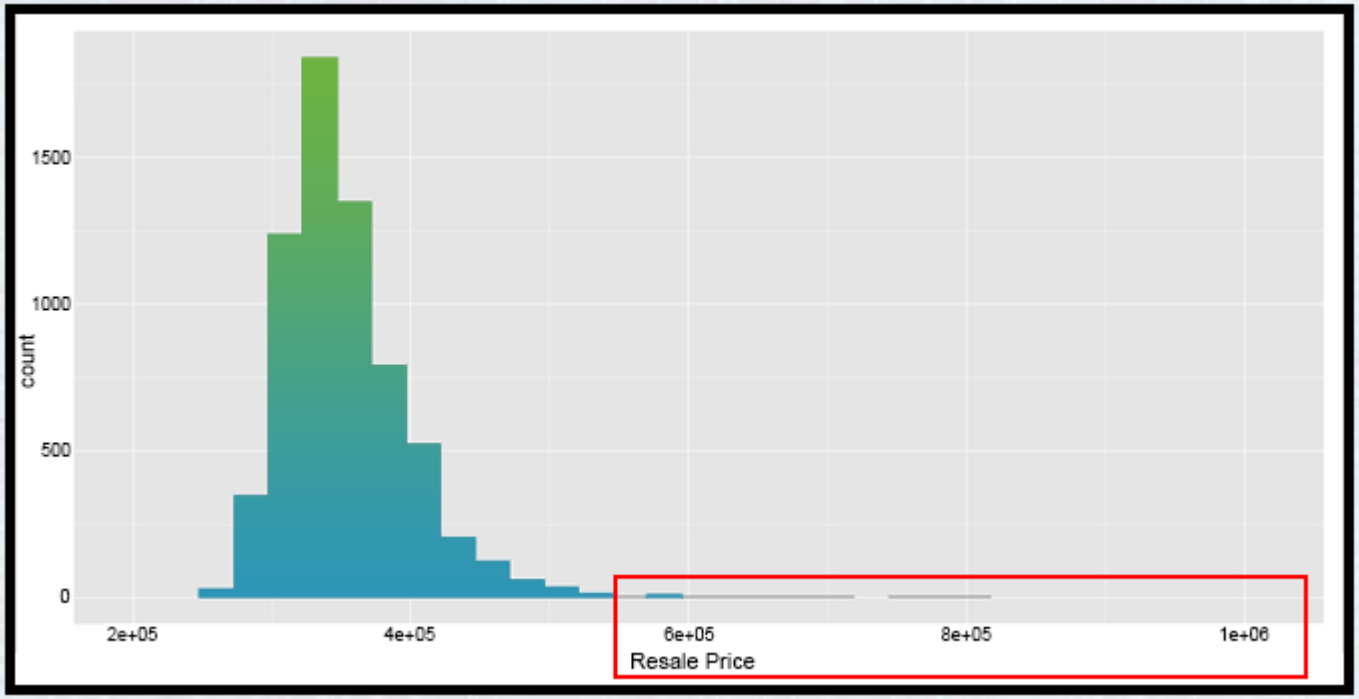
Note: The new features DaysElapsedSinceTransaction and Age will be taken forward, while the features LeaseCommenceDate and Approval Date will be dropped from further analysis.

Linear Regression

Resale Prices: Outliers

Liam looks at the plot of the Resale Prices. It reveals the distribution of numerical data. He could see a right skew in the data—a long tail to the right. It indicates there are apartments which have been sold at an exceptionally high price, an outlier in the current context.

Refer to the image here to view the plotting of the Resale Prices.

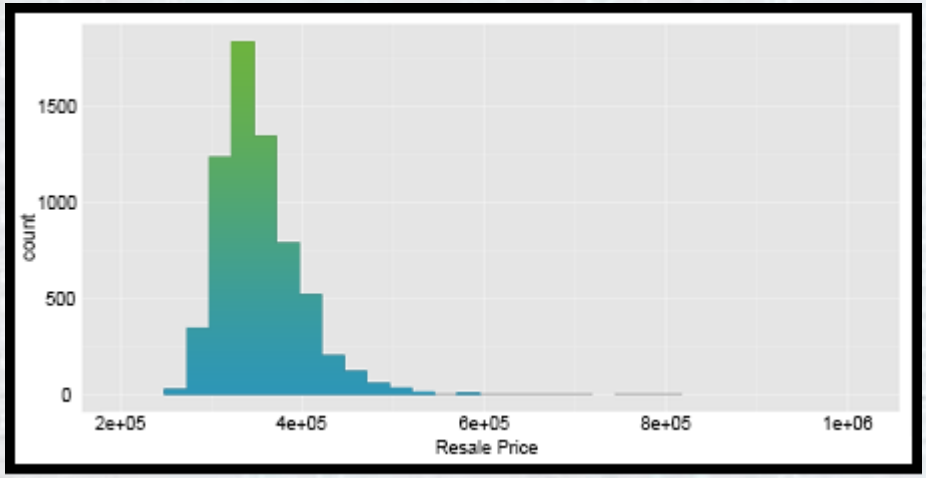


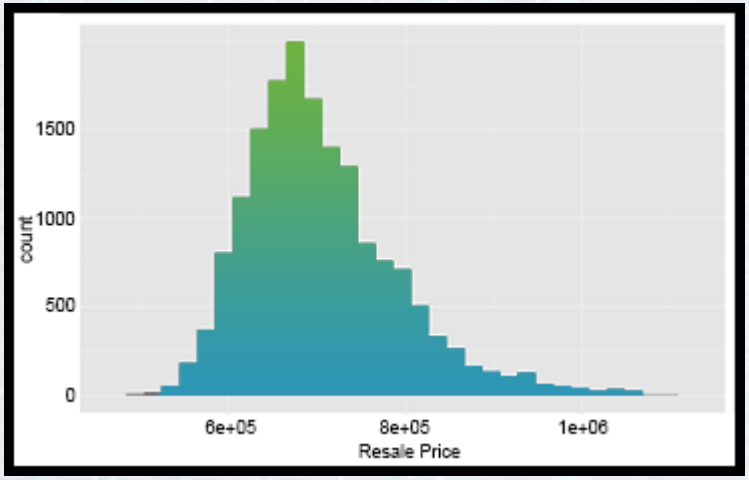
Linear Regression

Resale Prices: Removing Outliers

Since Liam's objective is to create a predictive model for the normal population, he removes the extreme outliers. This removal ensures that the outliers do not affect the current modeling exercise.

The outliers comprise 1% of the data. Therefore, only 72 data points are removed by Liam. He now re-plots the distribution of resale price.



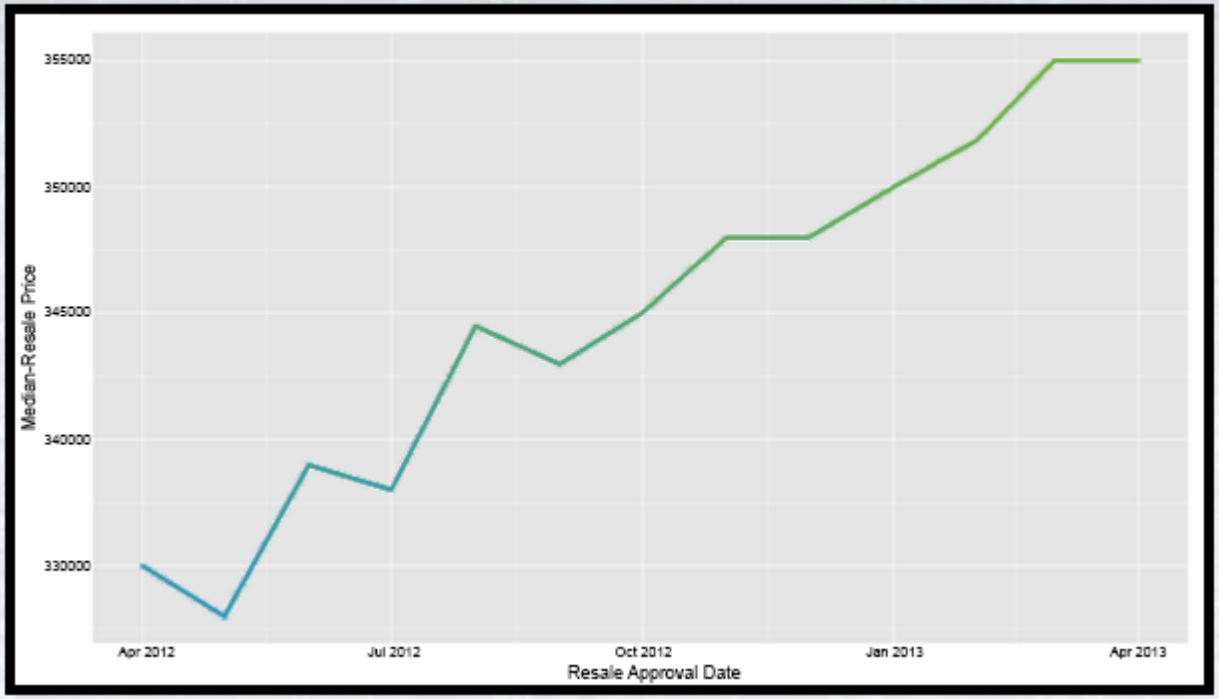


Linear Regression

Analysis of Resale Prices across the Timeline

Liam plots the median of Resale Price against the Resale Approval Date. Take a look at the graph here to view what Liam finds out.

As you can see, there is an increasing linear path with minor fluctuations across the timeline. Such a graph suggests that the input data is suitable for Linear Regression.



Linear Regression

Feature Scaling

While analyzing the data, Liam comes across another challenge—some features available in the dataset are in different scales.

For example, the SQMS reflects the size of an apartment and is in the scale of 100s. On the other hand, the bedrooms is in the order of 1, 2, and 3.

Liam decides to scale the features using Mean Normalization. It is one of the techniques through which the Feature Scaling can be done.

The formula of Mean Normalization is given below:

$$\frac{x - \text{Mean}(x)}{\text{StandardDeviation}(x)}$$

Linear Regression

Categorical Data Points

In the meanwhile, Liam has identified the three key categorical data points:

- Area
- Floor
- FlatType

He would transform the categorical data points into an Indicator Column or a Dummy Column to ensure no artificial bias is introduced.

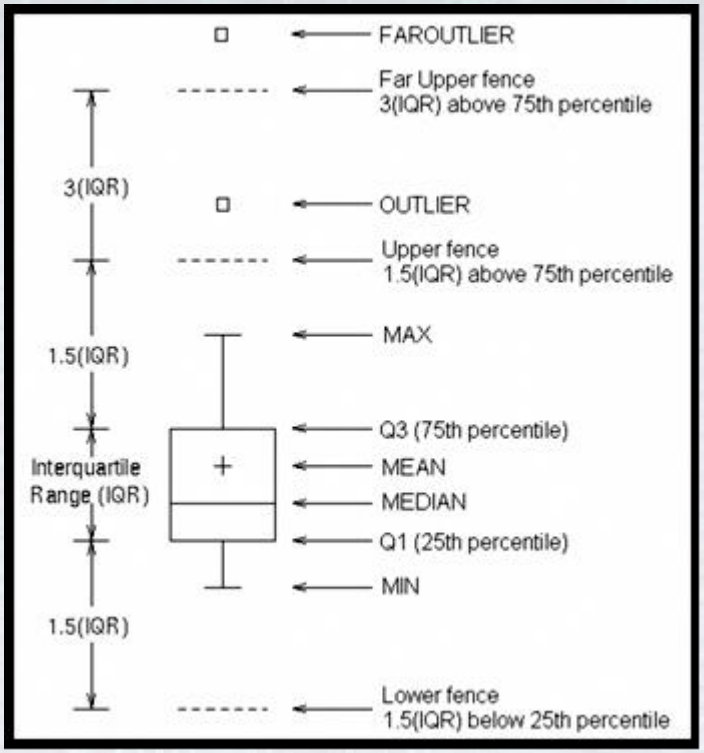
Queenstown	Marine Parade	Kallang-Whampoa	Clementi	Central Area	Bukit Merah	01 to 05	06 to 10	11 to 15	Improved	Terrace	Model A	New Generation	Standard
1	0	0	0	0	0	1	0	0	1	0	0	0	0
0	1	0	0	0	0	0	1	0	0	0	1	0	0

Linear Regression

Data Analysis: Box Plots

Liam's next task is to analyze the Resale Price by Area through Box Plot. It is a convenient way of graphically depicting groups of numerical data through their quartiles. This will help him identify the outliers.

However, Leslie wants to learn more about Box Plot. Liam shows him a diagram and starts explaining it in detail.



Linear Regression

Data Analysis: Box Plots

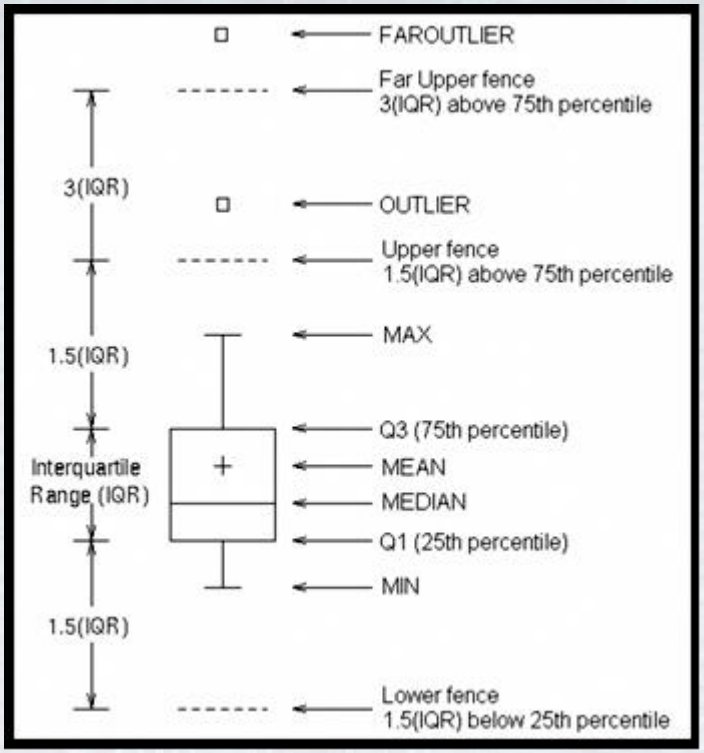
Box Plot

The box plot is a standardized way of displaying the distribution of data based on the five number summaries:

- Minimum
- First quartile
- Median
- Third quartile
- Maximum

A quartile is a group of values that divide a data set into quarters, or groups of four.

The Inter Quartile Range (IQR) is Q3–Q1.

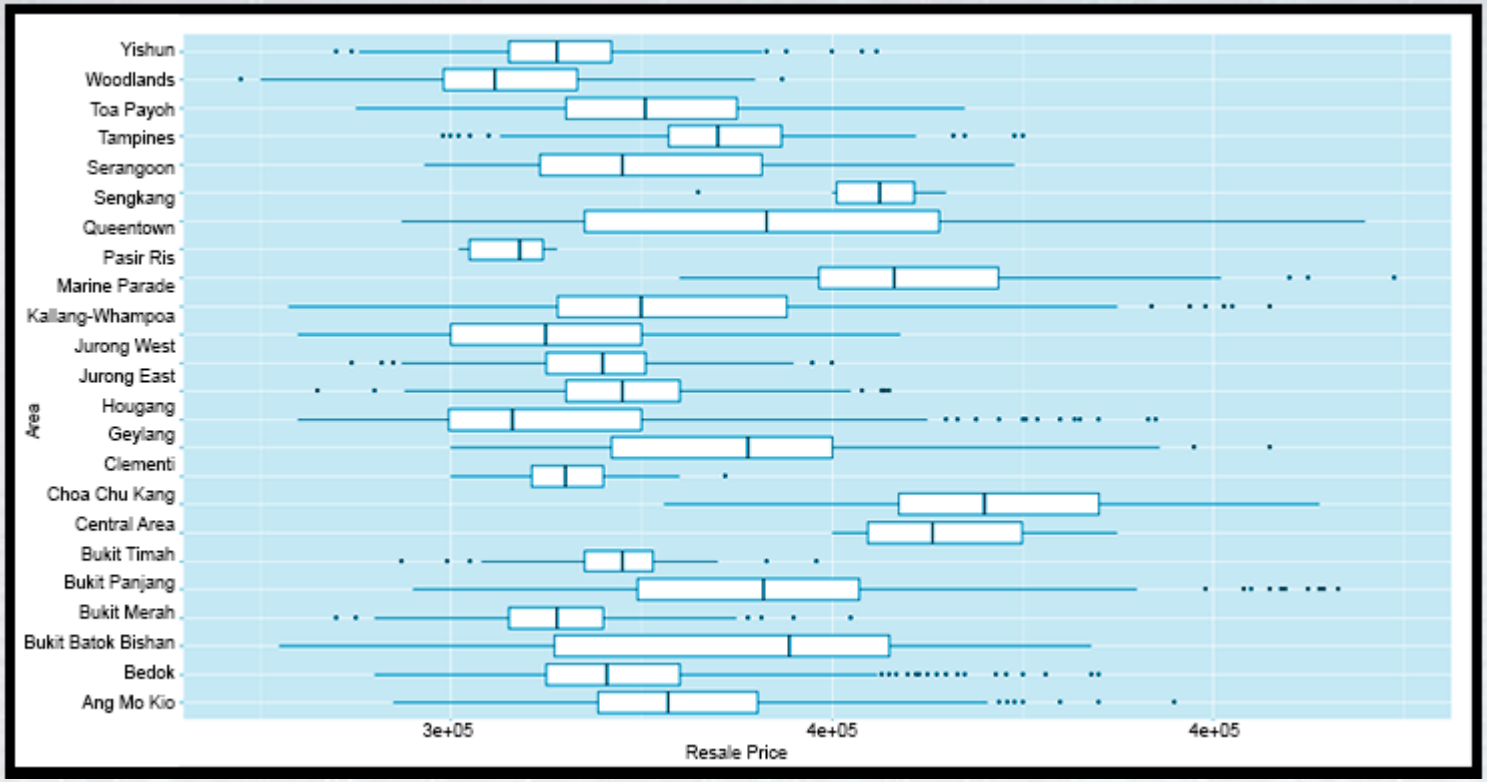


Linear Regression

Data Analysis: Box Plots

As Liam analyzes the Resale Price by Area through a Box Plot, he comes up with the following graph. It shows the distribution of the Resale Price across the areas in Singapore.

Refer to the image provided here to have a look at the graph that Liam developed.



Cognizant

Linear Regression

Check Your Understanding

State whether the sentence is true or false.

In Linear Regression, the outcome can have any one of the infinite number of possible values.

- True
- False

Linear Regression

Check Your Understanding

Correct Answer:

True

Point to Remember

In Linear Regression, the outcome can have any one of the infinite number of possible values.

Linear Regression

Input to the Algorithm

Liam has identified FlatType, Bedrooms, ResalePrice, and Age of the building (derived from LeaseCommenceDate) as the features to be used as an input to algorithm.

What are the additional features he needs to input into the algorithm?

- Area
- Address
- Floor
- SQMS

Linear Regression

Input to the Algorithm

Correct Answers:

- Area
- Floor
- SQMS

Point to Remember

Along with FlatType, Bedrooms, ResalePrice, and Age, the features Liam should use for input to the algorithm include:

- Area
- Floor
- SQMS

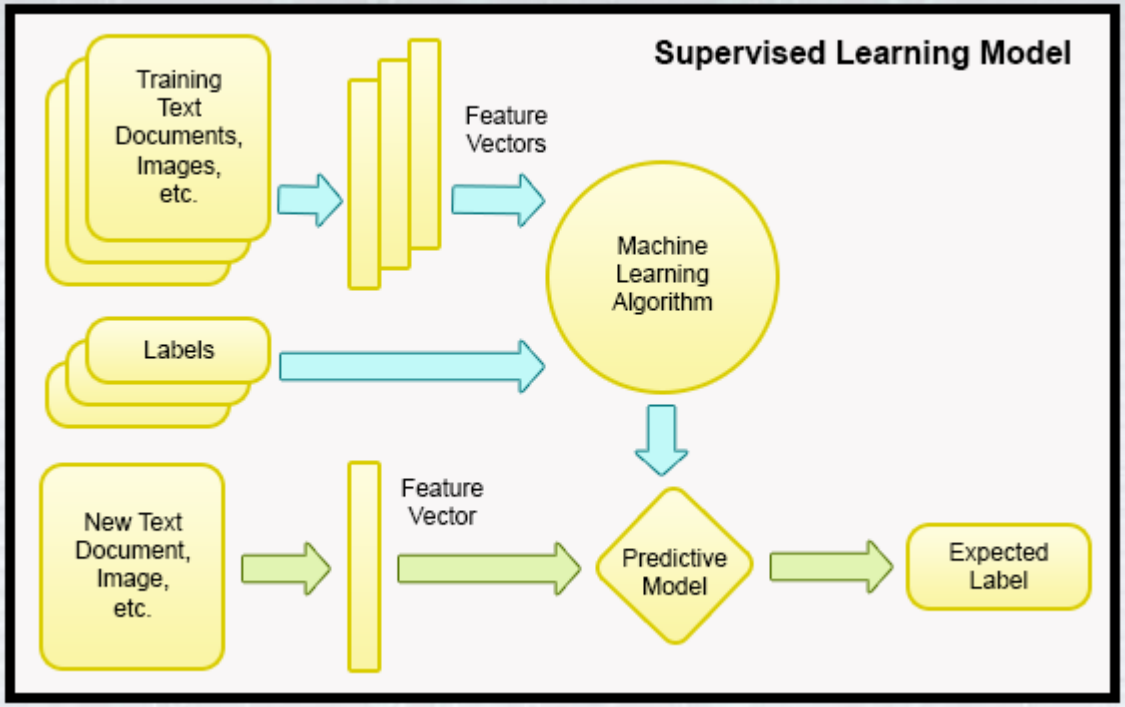
Linear Regression

Machine Learning Model

Liam is now going to enter the details into the Machine Learning model. He knows he has to customize the model as per the dataset he has got.

The current tools in the market such as Azure ML, SPSS, R, and Python offer out-of-the-box functions to apply diversified Linear Regression models on a given dataset. The dataset comprises:

- Features
- Response Variable



Linear Regression

Machine Learning Model

Labels

This is where the response variable should go. It should include the following:

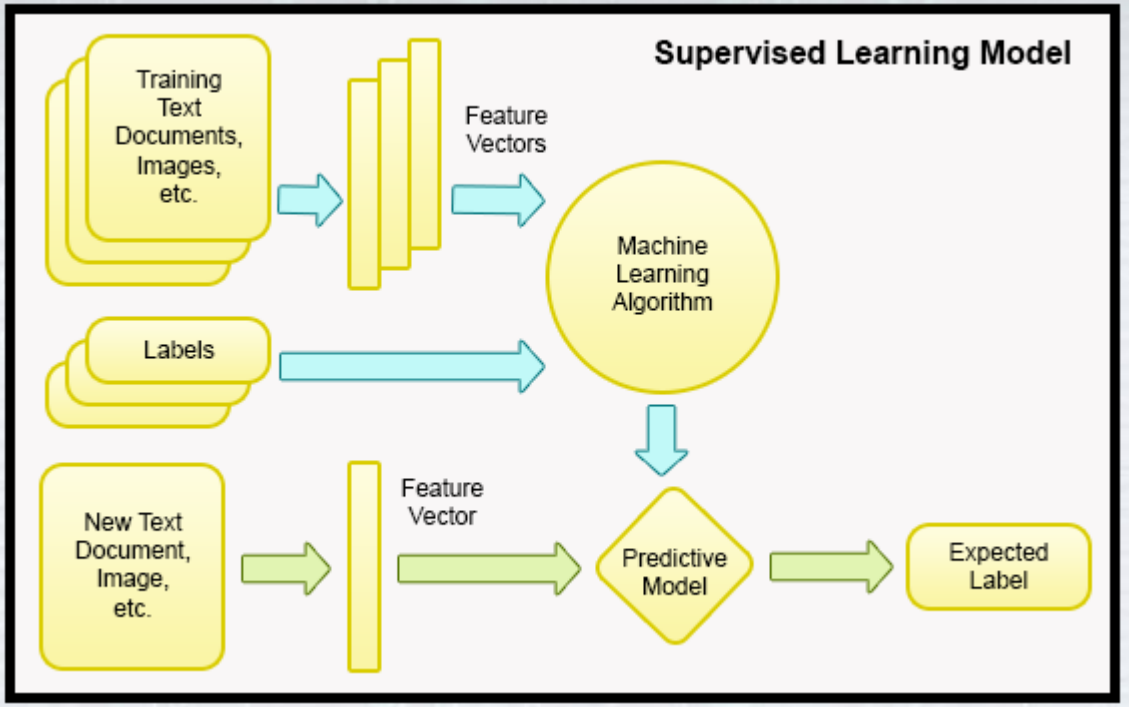
ResalePrice

Resale Price:
370000
361000
358000

Feature Vectors

This is where the features derived from the train dataset should go. It should include the following:

Area	Floor	SQMS	Flat Type
Bedok	16 to 20	65	Improved
Bukit Merah	01 to 05	74	New Generation
	11 to 15	60	Standard



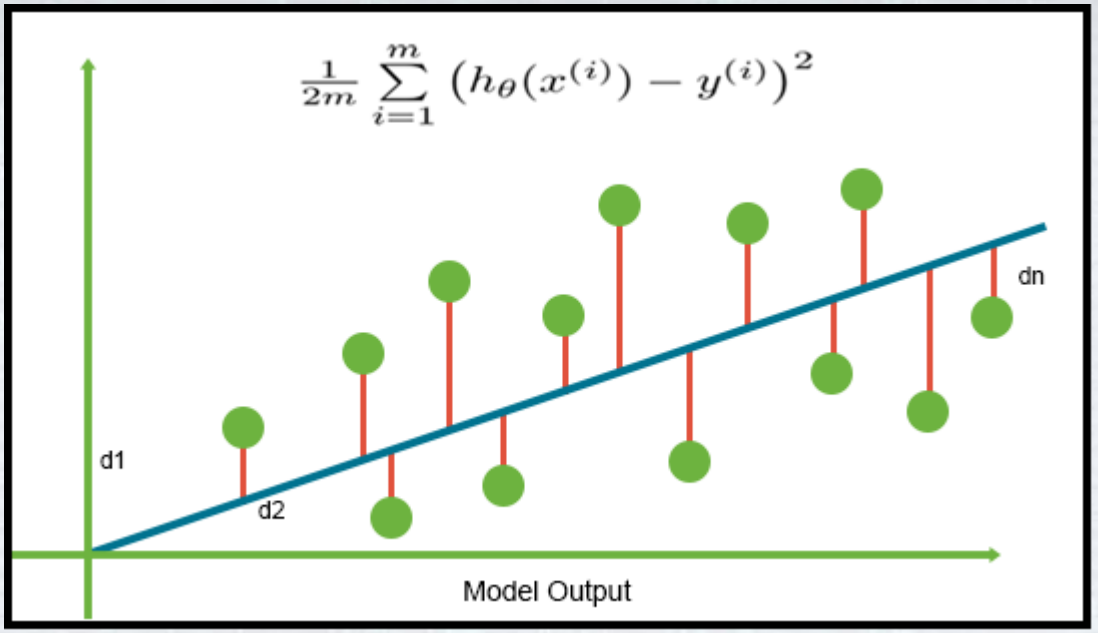
Linear Regression

Model Output

After customizing the algorithm model to suit the current requirement, Liam now checks the output. Refer to the graph provided here to view the output of the Machine Learning model.

While checking the model output, keep the following things in mind.

- In a mathematical perspective the Linear Regression is represented by a slope or linear function where a line passes through an optimal path.
- The Error is the difference between the Predicted and Actual Resale Prices.



Linear Regression

Error Handling and Feature Selection

Liam has got the model output now. However, he knows the importance of evaluating the output for error handling and feature selection.

There are two parameters on which a Linear Regression model is evaluated.

- Coefficient of Determination or R Square: The higher the value, the better the model
- Root Mean Squared Error or RMSE: The lower the Error, the better the model

The thumb rule is that the Coefficient of Determination should be high. The Coefficient of Determination is inversely proportional to RMSE.

The higher the Coefficient of Determination, the lower the RMSE.

Linear Regression

Machine Learning Process: Overview

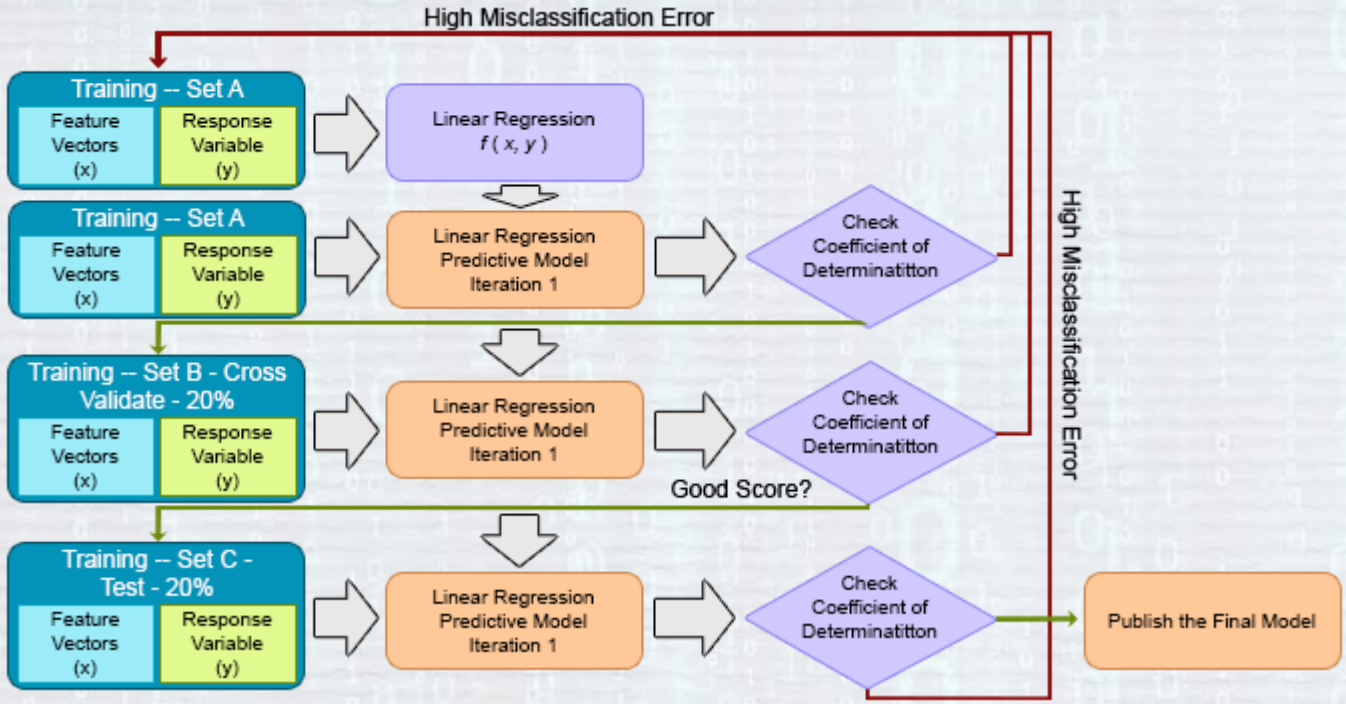
Now Liam decides to use the machine learning process to find out the results.
He has to do it in four phases. Let's take a look.

Phase 1

Phase 2

Phase 3

Phase 4



Cognizant

Linear Regression

Machine Learning Process: Overview

Now Liam decides to use the machine learning process to find out the results.
He has to do it in four phases. Let's take a look.

Phase 1

Phase 1:
The Set A of training data is passed as an input into the Linear Regression function.

Phase 2

Phase 2:
The resulting model is tested against the original Set A data.

Phase 3

Phase 3:
Linear Regression results are evaluated using Co-efficient of Determination and Root Mean Square Error. The model is then tested against the Set B of the training data.

Phase 4

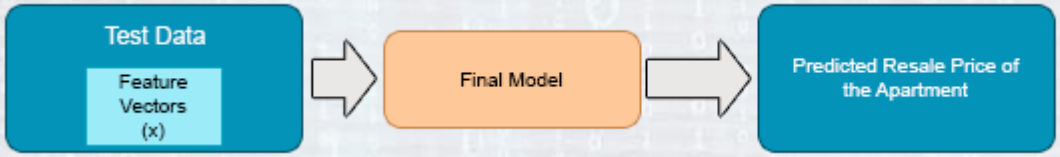
Phase 4:
Linear Regression results are evaluated using Co-efficient of determination and Root Mean Square Error. Then, the model is tested against the test data or the Set C of the Training Data.

Linear Regression

Machine Learning Process: Overview

For the last phase of the process, Liam modifies it slightly. He knows the Test Dataset does not have a Response Variable which is Resale Price in this context. In fact, he is trying to find it.

He modifies the last phase in the way mentioned below:



Note how the Test Data he entered consists only of Feature Vectors.

Linear Regression

Errors in a Predictive Model

While using the predictive model, Liam has to deal with the two types of errors that might be present in a predictive model: Bias Problem and Variance Problem.

Bias Problem

Variance Problem

Linear Regression

Errors in a Predictive Model

Bias Problem

It is an under fitting problem; and it does not respond well to the training data itself. The possible strategy for bias problem is to:

- Add more features or derived features based on the existing features.

Variance Problem

It is an over fitting problem, and fits well to the training data. However, the model does not fit to the cross validation or test data set. Possible strategies for variance error are to:

- Add more dataset to the training data.
- Reduce the number of features.

Linear Regression

Check Your Understanding

What are the possible strategies to counter variance problem?

- Add more dataset to the training data.
- Reduce the number of features.
- Add the number of features.
- Add derived features based on the existing features.

Linear Regression

Check Your Understanding

Correct Answers:

- Add more dataset to the training data.
- Reduce the number of features.

Point to Remember

Possible strategies to counter variance problem are to:

- Add more dataset to the training data.
- Reduce the number of features.

Linear Regression

Summary

Some of the key points of the course are mentioned below:

- Linear Regression is based on Linear Algebra. It inputs the features of the train dataset to deduce the outcome of the test dataset.
- While using the dataset, you have to analyze the data, identify the relevant features, plot their distributions, and convert them into indicator form, if applicable.
- It is a good strategy to plot the median of the model output. If the output shows increasing linear pattern, the data is suitable for Linear Regression.
- There are two parameters on which a Linear Regression model is evaluated: Coefficient of Determination or R Square, and Root Mean Squared Error, or RMSE.
- Error handling is an effective way to evaluate the model outcomes and ensure the prediction is accurate.

Cognizant

Thank You