**(ML) Section 4: Assignment 1 Main Page -Understanding How Supervised ML Models Work**

Main Learning Goal

The main goal of this unit is to understand how you train Machine Learning models to analyze data. This unit also looks at how different ML models can be used in order to and aid data-based decision making using financial information as a real-world example.

Focus Question: How can you train ML models to analyze data?

Background

As was explained earlier, supervised learning is a machine learning approach that’s defined by its use of labeled datasets. Supervised learning can be separated into two types of problems when data mining: **Classification**and **Regression**:

**Classification** **Problems**- use an algorithm to accurately assign test data into specific categories, such as separating apples from oranges. Or, in the real world, supervised learning algorithms can be used to classify spam in a separate folder from your inbox.    
Linear classifiers, support vector machines, decision trees and random forest are all common types of classification algorithms.

**Decision Tree**

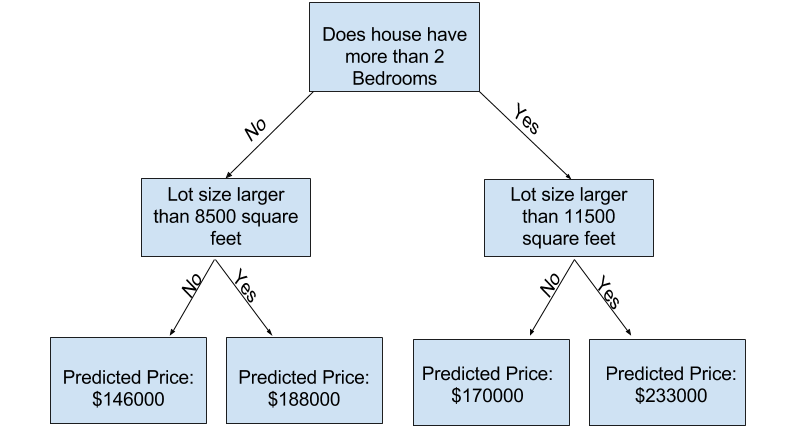
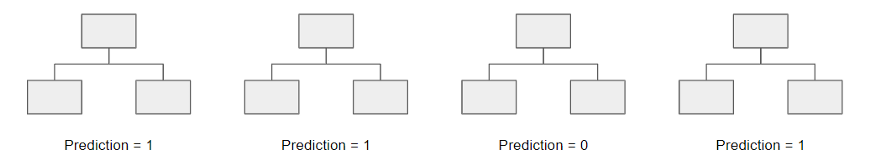


Image taken from Kaggle

**Decision trees** - a popular model used in operations research, strategic planning, and machine learning. Each square above is called a **node**, and the more nodes you have, the more accurate your decision tree will be (generally). The last nodes of the decision tree, where a decision is made, are called the **leaves** of the tree. Decision trees are intuitive and easy to build but fall short when it comes to accuracy.

**Random forests** - an [ensemble learning.](https://en.wikipedia.org/wiki/Ensemble_learning) technique that builds off of decision trees. Random forests involve creating multiple decision trees using [bootstrapped datasets.](https://machinelearningmastery.com/a-gentle-introduction-to-the-bootstrap-method/) of the original data and randomly selecting a subset of variables at each step of the decision tree. The model then selects the mode of all of the predictions of each decision tree. What’s the point of this? By relying on a “majority wins” model, it reduces the risk of error from an individual tree.



For example, if you created one decision tree, the third one, it would predict 0. But if you relied on the mode of all 4 decision trees, the predicted value would be 1. This is the power of random forests.

StatQuest does an amazing job walking through this in greater detail:

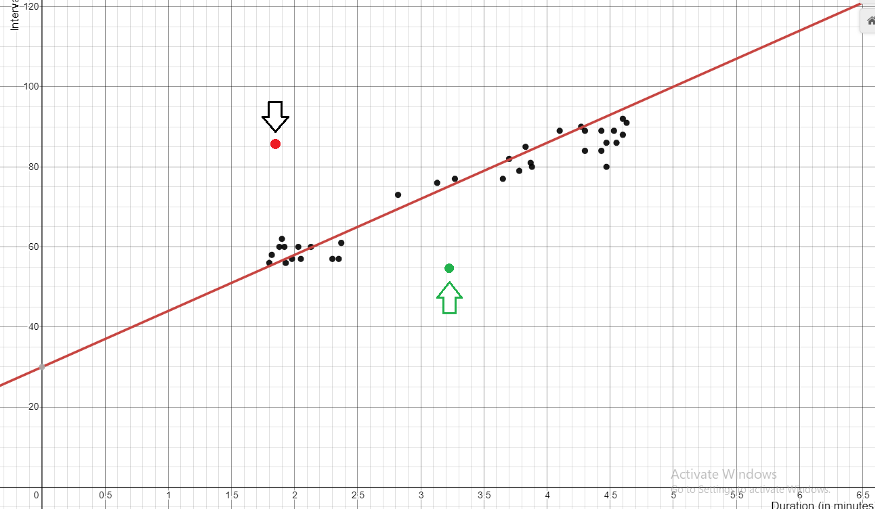
[StatQuest: Random Forests Part 1 - Building, Using and Evaluating - YouTube](https://www.youtube.com/watch?v=J4Wdy0Wc_xQ&t=13s)

**Regression**- A different supervised learning method that uses an algorithm to understand the relationship between dependent and independent variables. Regression models are helpful for predicting numerical values based on different data points, such as sales revenue projections for a given business. Some popular regression algorithms are linear regression, logistic regression and polynomial regression.

**Linear Regression** - one of the most widely used approaches used to model the relationship between two or more variables. It is widely used in science and engineering applications and may  be familiar to your students. It can be applied to things like sales forecasting for inventory planning, or to determine impact of greenhouse gases on global temperatures, to predicting crop yield based on rainfall, or predicting which teams will advance to playoffs.

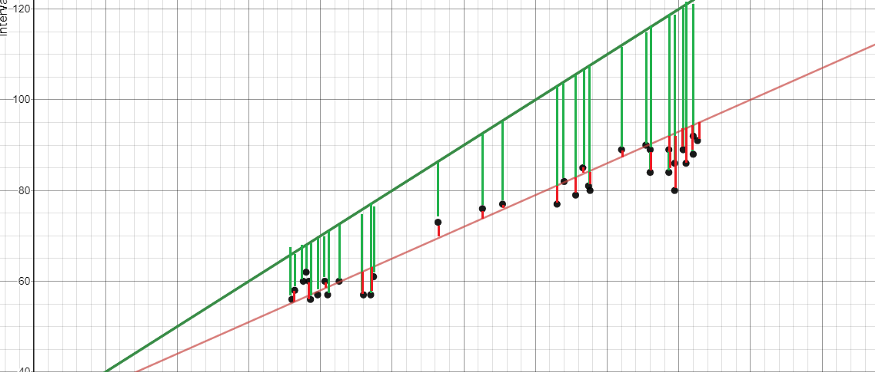
According to Wikipedia, linear regression is a linear approach to modeling the relationship between a dependent variable and one or more independent variables. In simpler terms, it is the ‘line of best fit’ that represents a dataset.

Below is an example of a line that best fits the data points. By creating a line of best fit, you can predict where future points may be and identify outliers. For example, assume that this graph represents the price of diamonds based on weight. If you look at the red dot, you can see that this particular diamond is overvalued because it costs much more given the same weight as other diamonds. Similarly, the green dot is undervalued because it costs much less than other diamonds with similar weights.



So how do you find the line of best fit? Let's find out.

How Simple Linear Regression works



You’re going to focus on **simple linear regression**. The line of best fit, or the equation that represents the data, is found by minimizing the squared distance between the points and the line of best fit, also called the **squared error**.

To give an example, there are two ‘line of best fits’ shown above, the red line and the green line. Notice how the error (the green lines between the line of best fit and the plots) is much greater than the red line. The goal of regression is to find an equation in which the sum of the errors is minimized.

If you want to know the math behind it, you can watch Khan Academy’s videos [hereLinks to an external site.](https://www.khanacademy.org/math/statistics-probability/describing-relationships-quantitative-data/more-on-regression/v/squared-error-of-regression-line" \t "_blank), where they find the partial derivatives of m and b.

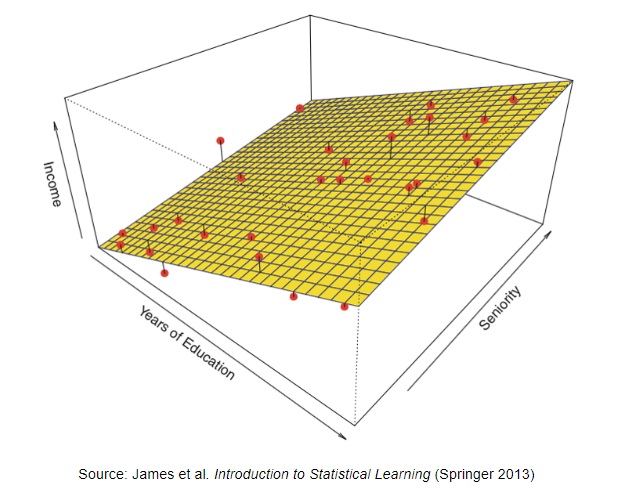
**Multiple Linear Regression**

Simple linear regression is useful when you want to find an equation that represents two variables, the independent variable (x) and the dependent variable (y). But what if you have many independent variables? For example, the price of a car is probably based on multiple factors, like its horsepower, the size of the car, and the value of the brand itself.

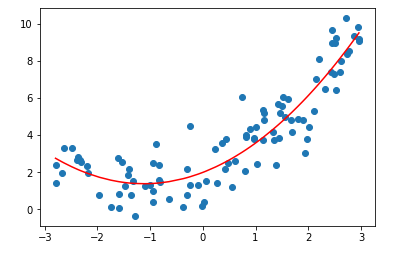
This is when multiple regression comes in. **Multiple regression** is used to explain the relationship between a dependent variable and more than one independent variable.

The image below shows a plot between income (y) and seniority and years of education (x). When there are two independent variables, a **plane of best fit** is found instead of a line of best fit.

**Polynomial Regression**



What if you had a set of data where its line of best fit is not linear (like the image below). This is when you would want to use **polynomial regression**. Using Wikipedia again, it’s defined as a form of regression analysis in which the relationship between the independent variable x and the dependent variable y are modeled as an nth degree polynomial in x. In simpler terms, it fits a non-linear relationship between x and y.



TO DO:

1. Assignment 1 - How do we train and test models?