Topic: Introduction

Data Rich Business problems

Assuming we have enough data to proceed with the analysis. Our next decision is to look at the outcome we're trying to predict and determine if it's a numeric outcome or a non numeric outcome. Numeric outcomes are those where the outcome is simply a number. Predicting the demand for electricity or the hourly temperature are both numeric outcomes. non numeric outcomes are those where we're trying to predict the category into which a case or customer falls, such as whether a customer will pay on time, pay late or default on a payment. Another example is whether or not an electronic device will fail before 1000 hours. We generally refer to models that predict a non numeric outcome as classification models. Those that Predict Numeric outcomes are known as regression models.

Key Points/Summary:

The text explains how to choose between two types of machine learning models based on the kind of outcome you're trying to predict. If you want to predict a number, like temperature or demand, you'd use a regression model. If you're trying to predict a category, like whether a customer will default on a payment, you'd use a classification model.

Data Poor Business Problems

You don't have enough usable data to solve the problem, and we need to set up an experiment to help us get the data we need. An experiment in the business context is usually referred to as an AB test. For example, you want to know how replacing an old product with a similar new product will impact sales. Since you don't have data for sales of the new product. You don't have data to do any modelling or prediction. To determine sales of the new product, you can set up an experiment. You can introduce the new product in just a few stores to help estimate the impact on sales if you introduce the product at all stores. Just taking any stores for your experiment won't work very well. Therefore, it's important to design your experiments so you can use the results at other stores. I'll go into more detail and of course focused entirely on a B testing

Key Points/Summary:

The passage describes using A/B testing to gather data when there's not enough existing information. It uses the example of wanting to predict sales of a new product. Since there's no sales data for the new product, an A/B test can be conducted. This involves introducing the new product in a limited set of stores to estimate its impact on overall sales if introduced everywhere. However, simply choosing any stores won't provide reliable results. The design of the experiment is crucial to ensure the results can be applied to other stores.

Introduction To Non-Numeric Models

Let's move on to classifications and the modelling types that are appropriate for non numeric target variables. Earlier, we learn the classifications or non numeric outcomes are those where we're trying to predict the category into which a case falls such as whether an electronic device will fail before 1000 hours or not. Another example is whether a customer will pay on time pay late or default on a payment. For example, you might classify the size of stores into large, medium or small categories. When modelling categorical variables, the number of possible outcomes is an important factor. If there are only two possible categorical outcomes, such as yes or no or true or false, then the variable can be described as binary. If there are more than two possible categorical outcomes, such as small, medium or large, or pay on time, pay late or default on a payment then the variable can be described as non binary. The important takeaway from this lesson is the ability to determine if you should use a classification model and whether it should be a binary model or non binary. Ben Burkholder will lead a course focused on classification models and they'll go into detail about these types of models.

Key Points/Summary:

This passage discusses how to identify the right type of classification model based on your target variable. Classification models deal with predicting categories, not numbers. Examples include predicting customer payment behavior (on-time, late, default) or electronic device failure.

The key takeaway is understanding how many categories your target variable has. If there are only two (yes/no, pass/fail), it's a binary classification problem. If there are more than two categories (payment options, clothing sizes), it's a non-binary classification problem. This distinction helps choose the appropriate classification model for your analysis. The passage mentions a follow-up course that will delve deeper into specific classification models.

Introduction To Numeric Models

Great job on the quiz. Now let's dive into numeric models by introducing target variables. Since the underlying math uses variables and equations, and we're effectively solving an equation for a variable that represents the outcome, we'll be using target variable to represent the outcome we're solving for. Alright, so now that we know how to determine if the target variable is numeric or non numeric, we need to investigate the target variable a little further. This will help us identify the most appropriate model of First, let's focus on numeric variables. At this point, we need to determine what type of numeric variable the target is. There are three types of numeric variables continuous time based and count. The first type of variable we'll discuss this continuous. A continuous variable is one that can take on all values in a range. For instance, your height can be measured down to many decimal places, we don't grow at even inch intervals. The second type of numeric variable is a time based variable. The time based numeric variable is one where you're trying to predict what will happen over time. This is often related to forecasting the third type of numeric variable is a count variable. Count variables are numbers that are discrete positive integers. They're called count numbers because they're used to analyse variables that you can count as modelling these types of variables is not common in business. We won't be covering this topic in this course. Now that we know the types of numeric variables, we can find the types of models that are appropriate for the target variable. If the target variable is a continuous variable, we can build continuous models to solve the business problem. If the target variable is time based, we can do time series analysis to solve the business problem. The important takeaway from this lesson is the ability to determine if

you should use a numeric model and whether it should be a continuous or time based model. I do want to take a moment to highlight the fact that there are many more models available to us. But the focus of this course is on some of the more commonly used models in subsequent courses will go into much more detail about the specific models.

Key Points/Summary:

This passage explains how to choose a model for numeric target variables in machine learning.

- **Target Variable:** The outcome you're predicting in your model, either numeric (numbers) or non-numeric (categories).
- **Types of Numeric Variables:**
    - **Continuous:** Can take any value within a range (e.g., height, weight). Use continuous models for these.
    - **Time-based:** Focuses on predicting future values over time (e.g., sales forecast). Use time series analysis for these.
    - **Count:** Whole numbers you can count (e.g., number of customers). Not covered as uncommon in business modeling.
- **Choosing a Model:** The type of numeric variable you have determines the model type:
    - Continuous variable - Use continuous models.
    - Time-based variable - Use time series analysis.

By understanding your target variable's type (numeric and continuous/time-based), you can choose the right model for your business problem. The course focuses on common models with more details coming in future lessons.

Descriptive

Lastly, let's discuss some basic descriptive statistics concepts. In short, descriptive statistics provides simple summaries of a data sample. Examples could be calculating an average GPA for applicants to a school, or calculating the batting average of a professional baseball player. In our electricity supply scenario, we can use descriptive statistics to calculate the average temperature per hour per day or per date. Some of the commonly used descriptive statistics are Mean, Median mode, standard deviation, and interquartile range. Udacity has an entire course dedicated to descriptive statistics, which I encourage you to look at if you're unfamiliar with these terms. And concepts.

Key Points/Summary:

The passage discusses basic descriptive statistics concepts used to summarize data. Descriptive statistics provide simple summaries of a dataset, like calculating an average GPA or temperature. Examples include calculating average temperature per hour for electricity supply. Some common descriptive statistics are mean, median, mode, standard deviation, and interquartile range.

Topic: Feature Selection

Variable Reduction Example

Let's look at an example of using variable reduction. Say we are analysing an employee survey where the employee has to rate with a response from one to seven the following questions. My boss treats me with consideration my boss gives me recognition when I do a good job. My boss consults me when there are decisions to be made about my work. Now in reality, these are all similar and could probably be categorized as employee satisfaction with supervisor an analyst could choose to leave them all in. But again, if they're already dealing with a number of other variables, it might make it simpler to just have a single variable that captures most of the variance in the response of these three variables. Again, much easier to communicate the concept of employee satisfaction than to identify each questions results separately. One negative aspect to variable reduction is that while it does reduce the number of variables in the analysis, depending on the method used in the data, it can make the interpretation of the model more difficult. If the variables you are combining aren't necessarily related, for example, population totals and income levels, the component variables that are created might be hard to explain. It's a good practice to try to use variable reduction when the variables are somewhat related. We see this in the example data we built out the exercise from the previous lesson. We have 12 months each of temperature snowfall, rain and precipitation data. I wouldn't want to use variable reduction methods to combine all 48 variables at once. But it would make sense to create component variables for each category of climate. This way snowfall is kept separate from temperature. So as an analyst you need to weigh the benefits you can get from reducing the number of variables you use in the clustering analysis with the all important aspect of needing to be able to explain the basis for the clusters

Key Points/Summary:

This passage explains variable reduction in the context of cluster analysis, a technique for grouping similar data points.

- **Variable Reduction:** Reducing the number of variables used in an analysis.
- **Benefits:**
    - Makes analysis simpler (fewer variables to manage).
    - Easier to communicate findings (e.g., employee satisfaction vs. 3 separate questions).
- **Drawbacks:**
    - Can make interpreting the model difficult if combined variables aren't related.
- **Good Practice:** Combine variables that are conceptually similar.
- **Example:** Employee survey questions about supervisor can be combined into a single "employee satisfaction" variable.
- **Important Consideration:** Balance the benefit of fewer variables with the need to explain the reasoning behind the clusters. Don't combine unrelated variables if it makes interpreting the clusters unclear.

Factor Analysis and PCA Overview

there are two common methods that can be used for variable reduction. These are factor analysis or more technically correct exploratory factor analysis and principal components analysis. Let's go through the high level concepts and differences between each of these methods. Let's start by referring back to the employee survey example. You will recall there were three questions that could be related to the concept of employee satisfaction was supervisor. Factor analysis assumes that there is a latent or hidden variable called a factor underlying these three questions. This factor is what basically causes the way these three questions get answered. In other words, how satisfied you are with your supervisor will determine how you answer those questions. So if you know the factor, which in this case is really the rating for employee satisfaction with supervisor, you will be actually able to predict what the answers to the three questions are. Principal Components Analysis, which is commonly referred to as PCA, on the other hand, makes no assumptions about any underlying causal model instead, PCA essentially summarizes all the variants within the total variables into fewer components and reduces the number of variables so what is the primary difference between the two methods? PCA analyses all of the variance within all of the variables selected, while factor analysis only analyses the variances that are shared or common within the variables another way to put it is that PCA focuses on accounting for the total variation or as factor analysis focuses on accounting for just the correlations between the variables. So factor analysis is trying to answer the question of what might be causing the responses and PCA is trying to answer how it can summarize the variables.

Key Points/Summary:

The passage explains two methods for variable reduction: Factor Analysis and Principal Component Analysis (PCA).

Factor Analysis:

- Assumes an underlying cause (factor) explains related variables (e.g., employee satisfaction explains answers to supervisor questions).
- Aims to find this factor to predict individual variable values.

Principal Component Analysis (PCA):

- Makes no assumptions about underlying causes.
- Focuses on summarizing the total variation across variables into fewer components.

Key Difference:

- Factor Analysis: Analyzes shared variance between variables (correlations) to identify potential causes.
- PCA: Analyzes all the variance in all variables for efficient data reduction.

In simpler terms:

- Factor Analysis: Explains "why" variables are related.
- PCA: Explains "how" to summarize variables with minimal information loss.

Factor Analysis and PCA Continued

So how do you decide whether to use factor analysis or PCA? Factor analysis is best used when you're trying to understand the underlying factor or factors from a common set of variables with factor analysis, because you're estimating the underlying cause behind the variables. It becomes easier to conceptualize. When you have a goal to try to explain the correlations between the variables. Factor analysis works best. For example, it works well with social science analyses such as attitudinal studies or surveys, where you're trying to determine the underlying basis for the responses. So that's when you would choose factor analysis. But when would you use PCA? If the goal is to explain as much of the total variance as possible, where you want the total picture of the data, PCA is the choice but as mentioned earlier, because the components you end up with are more of abstractions of all of the data, those components may not be easy to interpret, or map out to an easy to understand concept. With PCA, the balance becomes having a reduced set of variables that are easier to work with, and yet being able to use them to tell the story to the business. That isn't necessarily easy, but we will show how it could work in our example exercise. Of course, as we shall see in the following exercise. It doesn't have to be an all or none approach to variable reduction. It may work well to have a hybrid where some of the variables are consolidated into components and the others are left alone.

Key Points/Summary:

The passage discusses how to choose between Factor Analysis and Principal Component Analysis (PCA) for variable reduction.

- **Factor Analysis:** Best when you want to understand the underlying cause(s) explaining related variables.
- **Use cases:**
    - Makes underlying causes easier to conceptualize (e.g., social science studies).
    - Explains the correlations between variables.
- **Principal Component Analysis (PCA):** Best when you want to capture the overall picture of the data by explaining as much of the total variance as possible.
- **Drawback:** The resulting components might be complex and difficult to interpret in a business context.
- **Choosing the right method:**
    - Factor Analysis - Understand the "why" behind related variables.
    - PCA - Reduce data complexity while capturing the overall picture.

The passage also acknowledges that a hybrid approach can be useful, where some variables are combined using PCA and others remain separate.

Topic: Linear Regression

Introduction to Linear Regression

Imagine we have the data displayed in a scatterplot. It appears that we have a linear relationship between the number of employees and the number of tickets. The relationship appears to be linear since it seems like you can draw a straight line through the data. If we know the equation of the line, we can predict values for tickets given a certain number of employees. The simple equation for a line is y equals mx plus b, where m is the slope of the line and b is the y intercept points. In our example, y is the average number of tickets and can be referred to as the target variable. So this is the variable we're trying to predict x, or the number of employees is the predictor variable, since we're trying to predict the number of tickets based upon the number of employees to determine the equation of the line we need to calculate the slope and the y intercept. With the equation, we can draw a straight line through our data. In this example, we'll use Google Sheets, but the same functionality exists in Microsoft Excel.

Key Points/Summary:

The passage describes how to find a relationship between two variables using a scatterplot. If the data points appear to follow a straight line, there's a linear relationship between the variables. We can then use the equation of that line (y = mx + b) to predict the value of one variable (target variable) based on the other variable (predictor variable). The passage mentions using spreadsheet software to calculate the slope (m) and y-intercept (b) to arrive at the equation of the best fit line.

Using Google Sheets To Calculate A Linear Equation

Using Google Sheets, let's determine the formula for a line with the x and y values that we have, where x is the number of employees, and y is the average number of tickets. First, to calculate the slope of the line, we use the slope function in Google Sheets. Note that all the functions used in this lesson are the same as in Microsoft Excel. The formula calculates the slope of 0.1833. Therefore, for every additional employee, the average number of tickets increases by 0.1833. Next, let's calculate the y intercept. For this we use the intercept function. The y intercept is minus 11.055. Now we know the equation of our line y equals 0.1833x minus 11.055.

Key Points/Summary:

This summary explains how to find the equation for a line representing the relationship between number of employees (x) and average number of tickets (y) using Google Sheets (or Microsoft Excel).

Here's a breakdown:

1. **Slope**: It's calculated using the SLOPE function, indicating an increase of 0.1833 in average tickets for every additional employee.
2. **Y-Intercept**: The INTERCEPT function reveals a y-intercept of -11.055.
3. **Equation**: Combining slope and intercept, the equation becomes y = 0.1833x - 11.055.

Linear Regression Validation

Now let's take a moment to revisit the problem solving framework that we started with. Now that we've perform the analysis and run the linear regression model, we need to validate the results of the model. In other words, is there a way to measure how good the model is, or in this case, is the linear expression we calculated a good bit of our data. Using the Correl function, we can calculate the correlation coefficient of the dataset or the variable r. The range of R is from minus one to plus one. The closer r is to plus or minus one the better correlation between X and Y. To learn more about correlation coefficients, see the link in the resources below. In our example, the value of r is 0.987, indicating a strong correlation. While strong correlation is good. We really want to know how well the data fits our line. Fortunately, you can get a sense of how good the formula is at approximating the data by calculating the coefficient and determination or r squared. R squared is the coefficient between zero and one r squared is interpreted as the percent of variance in observations that is explained by the model. An R squared value close to one would mean that nearly all variants in the target variable is explained by the model. An R squared value close to zero, would mean that nearly none of the variance in the target variable is explained by the model. An R squared value greater than 0.7 is considered to be a strong model. In practice, an R squared value of point five or greater is usually pretty good. An R squared value less than point three is generally agreed to not be useful.

Key Points/Summary:

This passage discusses how to assess the quality of a linear regression model after you've obtained the equation (linear expression). Here's a breakdown of the key points:

- **Validation**: We need to check if the model accurately captures the data's trend.
- **Correlation Coefficient (r)**: Calculated using the CORREL function, it measures the strength of the relationship between variables (x and y in this case). Its range goes from -1 to +1. The closer r is to +1 or -1, the stronger the correlation (positive or negative).
- **Coefficient of Determination (R-squared)**: This value (between 0 and 1) indicates the proportion of variance in the target variable (y) explained by the model.
    - Close to 1: Excellent fit - model explains most of the variation in y.
    - Close to 0: Poor fit - model explains little of the variation in y.
    - 0.7: Strong model - explains a significant portion of the variation.
    - = 0.5: Generally good model.
    - < 0.3: Not a useful model.

In summary, the passage provides ways to evaluate how well a linear regression model fits the data.

Validation in Google Sheets

We can square r or just use a function in Google Sheets RSQ to calculate the value of r squared, squaring the value of R gives us a value of 0.9744. Remember, the closer the data points are to the line, the closer the value of r squared will be to one if the value is closer to zero than the correlation is weak. In our example, the equation of the line is a great representation of the data that can be used to predict average values for the number of tickets. If we had a new customer with 600 employees, we could estimate the average number of tickets as y = 0.1833 x 600 - 11.0548, which equals 98.92, or approximately 100 tickets per week on average. We just created our first simple predictive model using linear regression and validated the result

Key Points/Summary:

This passage builds upon the previous one, explaining how to interpret the R-squared value and use the model for prediction. Here's the key takeaway:

- **R-squared (0.9744)**: Very close to 1 (excellent fit), indicating the model explains a high proportion of the variation in average tickets (y) based on the number of employees (x).
- **Prediction**: Given the model equation (y = 0.1833x - 11.055), we can estimate the average number of tickets for a new case (e.g., 600 employees). Plugging this value into the equation yields an estimated average of 100 tickets per week.

In essence, this passage highlights the successful creation of a simple linear regression model to predict average ticket values based on employee count.

Introduction To Multiple Linear Regression

What if we have more data available, can we determine if the additional data will result in a better prediction. We also have some data that shows the contract value per client against the average number of tickets. Using Google Sheets, we can plot the data with a linear equation and value for r squared. In this case, we have a linear equation with an R squared value of 0.785. This new data is similar to the original data when comparing values of r squared. But what if we could use both sets of data to develop a better predictor let's introduce multiple linear regression, which builds upon the simple linear regression by using more data to strengthen the correlation coefficient

Key Points/Summary:

This passage introduces the concept of multiple linear regression and its potential benefits. Here's a breakdown:

- **Additional Data**: The passage raises the question of whether including more data can improve prediction accuracy.
- **New Data**: It introduces a new dataset - contract value per client vs. average tickets - with its own linear regression model and an R-squared of 0.785, indicating a decent fit.
- **Multiple Linear Regression**: This technique expands on simple linear regression by incorporating multiple variables (e.g., both number of employees and contract value) to potentially achieve a stronger correlation coefficient (r). This suggests that using both sets of data in a multiple linear regression model might lead to a better prediction for average ticket values.

Multiple Linear Regression

We started with the simple equation of a line y equals mx plus b, where m is the coefficient of the variable and b is the y intercept. If you have more data that you believe has a linear relationship, you can expand that equation to y equals b0 + b1x2 + b2x3 + b3x3, and so on. In this equation, y is still the target variable. B0 is the intercept or the baseline value. b1 b2 b3 represent the coefficients of the variables, x1, x2, x3 and so on. The linear regression will find values for b0, b1, b2, b3. And these values represent the relationship we observe between the predictor variables and the target variable. This model is referred to as linear as the expression is based on a linear combination, an expression constructed from a set of terms by multiplying each term by a constant and then adding the results. With this dataset, our equation would look like this Y equals b0 + b1x1 + b2x2. Using Microsoft Excel or Google Sheets, we can calculate the values of the coefficients and check the correlation coefficients to see if our multiple linear model is a better predictor.

Key Points/Summary:

This passage explains how to perform multiple linear regression using a formula. Here's a breakdown:

- **Expanding the Formula**: It starts with the simple linear equation (y = mx + b) and expands it to accommodate multiple predictor variables (x1, x2, etc.).
- **New Notation**:
    - y: Target variable (same as before)
    - b0: Intercept (baseline value)
    - b1, b2, etc.: Coefficients for each predictor variable (x1, x2, etc.)
- **Finding Coefficients**: Linear regression calculates these coefficients (b0, b1, b2) to represent the relationship between the target variable and the predictor variables.
- **Model Type**: This remains a linear model because the equation is based on a linear combination of terms.
- **Example Equation**: The passage provides an example with two predictor variables (x1 and x2).
- **Software Tools**: It mentions using spreadsheet software (Excel or Sheets) to calculate the coefficients and assess the model's effectiveness using correlation coefficients.

Overall, this passage focuses on the mechanics of building a multiple linear regression model using a formula and software tools.

Multiple Linear Regression With Excel

Next, we're going to demonstrate an example of multiple linear regression with a data set that includes the average number of tickets, the number of employees and the value of the contract. For this example, we're going to use Microsoft Excel. You can perform multiple linear regression in Google Sheets with the linest function, but Excel provides a richer output. Note that to perform this next step, you need to have the Analysis Tool pack add in active in Microsoft Excel. To start the analysis. Go to the Data menu and select data analysis. Highlight regression and click OK. The input y range should be the range of your target variable, in this case, the average number of tickets. The input x range should be the range of data of your predictor variables, in this case, the number of employees and the value of the contract. Clicking OK we'll run the model and provide an output. We can quickly see our coefficients of our linear equation, which results in an equation of y equals minus 24.2667 plus 0.1019 x1 plus 0.00067 x2 where x1 is the number of employees and x2 is the value of the contract.

Key Points/Summary:

This passage dives into a practical example of performing multiple linear regression with Excel. Here's a summary:

- **Software**: Microsoft Excel is used for this demonstration (alternative: Google Sheets with linest function).
- **Analysis ToolPak**: This Excel add-in is required for the regression analysis.
- **Data Analysis Steps**:
    1. Go to the "Data" menu and select "Data Analysis."
    2. Choose "Regression" and click "OK."
    3. Specify the target variable range (average number of tickets) in "Input Y Range."
    4. Specify the predictor variables' range (number of employees & contract value) in "Input X Range."
    5. Click "OK" to run the model and get the output.
- **Results**: The passage highlights the obtained coefficients, allowing you to build the multiple linear regression equation:
    - y = -24.2667 + 0.1019x1 + 0.00067x2
        - y: Average number of tickets (target variable)
        - x1: Number of employees (predictor 1)
        - x2: Contract value (predictor 2)

Topic: Logistic Regression

Logistic Regression

Logistic regression is one of the most basic forms of regression modelling it's part of the family of generalised linear models or GLM for short, which basically means that the formula is very similar to that of a linear regression. However, since the target variable is in categories, instead of a continuous numeric variable, the target variable has to be modified to fit this GLM formula. Let's go back and revisit that formula from lesson one for a moment. Patrick presented the linear regression formula that you see on the screen here. As a reminder in this equation, y is the target variable. Beta naught is the intercept or the baseline value, and b1 or beta 1, beta 2, beta 3 all the way to beta n represents the coefficients for the different variables, x1, x2, x3 all the way to xn as well. So let's compare this formula to one of the logistic regression formulas. And yep there are actually more than one of them. Let's first focus on the right hand side of the equation. The right hand side of the equation is actually identical to that of a linear regression, where again, beta naught represents the intercept or the baseline value, beta 1, beta 2 beta 3 all the way up to beta n again, represents the coefficients of the variables x1, x2, x3 again, all the way to xn. Now what is going on the left hand side of the equation? Well, don't be scared. This is clearly where the logistic regression separates from that of a linear regression. But actually is conceptually very similar. What we're trying to do is find a probability of P which represents an outcome, like in our example, p represents the probability of yes, they will redeem the natural log is also in this equation to deal with the binary nature of this problem.

Key Points/Summary:

This passage compares linear regression with logistic regression. Here's a breakdown of the key points:

- **Logistic Regression**: It's a fundamental regression model type and belongs to the Generalized Linear Models (GLMs) family.
- **GLMs**: Their formulas resemble linear regression, but cater to categorical target variables.
- **Linear Regression Formula**: The passage revisits the formula ($y = \beta_o + \beta_1x_1 + \beta_2x_2 + ... + \beta_nx_n$) where:
    - y: Target variable (continuous numeric)
    - β_o: Intercept (baseline value)
    - β_1 to β_n: Coefficients for predictor variables (x_1 to x_n)
- **Logistic Regression Formula**: While the right side with coefficients (β) and variables (x) is similar to linear regression, the left side differs conceptually.
- **Left Side of Logistic Regression Formula**: It involves the natural logarithm (ln) to transform the target variable (represented by P) into a probability. In the example, P signifies the probability of a "yes" outcome (e.g., customer redeeming a product).

In essence, the passage highlights that logistic regression adapts the linear regression formula to handle situations where the target variable is a probability instead of a continuous numerical value.