Data Science

A Comprehensive Summary

Salvin Chowdhury

"Data science, as it's practiced, is a blend of Red-Bull-fueled hacking and espresso-inspired statistics. ... Data science is the civil engineering of data. Its acolytes possess a practical knowledge of tools and materials, coupled with a theoretical understanding of what's possible."

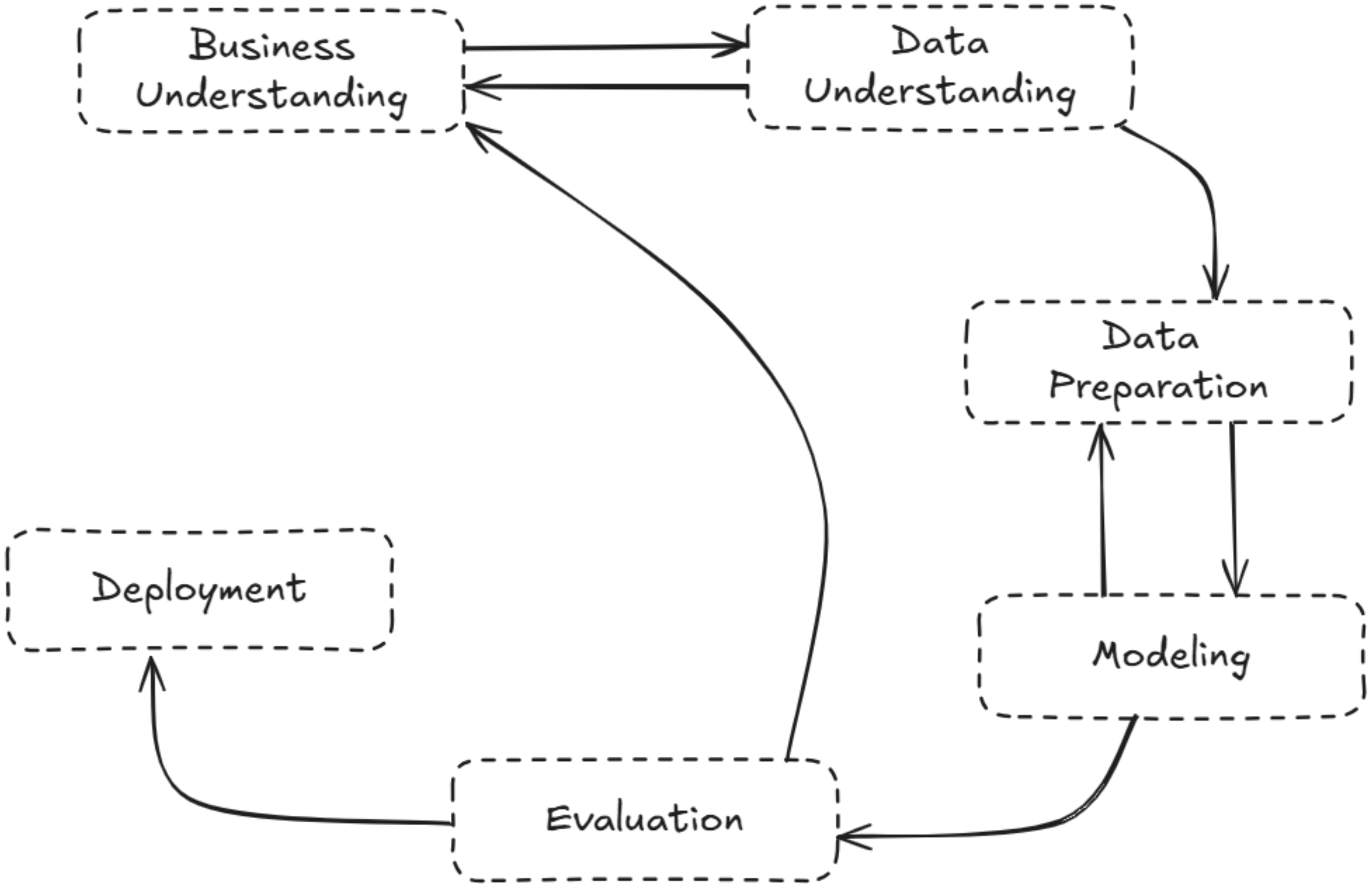
Mike Driscoll

The Process

Data Science Process

Data Driven Process

In Data Science, we start with a dataset, and explore it to identify the patterns. When we discover such patterns, we ask questions and then form the hypothesis. We then use the data to answer the hypothesis, else we design a new experiment.



Data Science Process

Introduction to CRISP-DM

CRISP-DM stands for Cross Industry Standard Process for Data Mining. It provides a framework for data mining projects and is organized into six phases

1 | Business Understanding

In this section, we focus on understanding the objectives and requirements of the project. We understand what the objectives and requirements of the project is. We define what success looks like from a technical data mining perspective and produce a project plan.

2 | Data Understanding

In this section, we focus on identifying, collecting, and analyzing the datasets that helps to accomplish the project goal. We collect the data, describe it, visualize and identify relationships and verify the data quality.

3 | Data Preparation

In this section, we select the required data, clean & construct the data, integrate any other data set that we might find and then format the data as deemed necessary.

Data Science Process

The Data Cleaning Steps

When it comes to data cleaning, we need to gather data, clean data, explore data, validate hypothesis and communicate the results. To do this, we do the following steps:

- Parse the file and convert it into consistent representations
- Convert values into right types, right units and remove any outlier records

Any next steps would be to clean up the data and organize it into a table.

Understanding the Data

We look at the first few records of the dataset, and use methods from Pandas to get information about the columns and get any descriptive summary statistics.

Data Cleaning Steps

There are other data cleaning steps that we need to perform such as:

- Converting features into a single representations, such as using abbreviations and etc.
- Determine the data types of each feature by using .astype(), such as categorical or numerical
- For any string data, see if they can be broken up any further if they're found to be complex
- All values in a feature must follow the same units

Data Science Process

Errors & Artifacts

When it comes to missing values, we need to look at the cause behind them. Some missing values may be meaningful, otherwise they can be imputed. We can remove any records with missing values or determine which imputation strategy makes the most sense (safest is mean or mode).

Data Types

Different Data Types

Categorical | Nominal vs Ordinal

Nominal data is the qualitative classification of objects by names (gender, nationality) and have no order information. Ordinal data has a natural ordering that measures rank. It has no quantitative value (education level, class).

Numerical | Discrete, Interval & Ratio

Discrete is the constant difference between units on a scale. It can be counted (age). Interval is the constant difference between units on a scale but has no zero point (temperature). Ratio is the constant difference between units on a scale and has a zero point (height).

Descriptive & Inferential Statistics

Descriptive statistics is used to summarize the information in a given sample, and inferential statistics is used to learn about the population that the sample comes from.

Central Tendency & Measures of Variability

Some measures of central tendency is the mode, median, mean & midrange. Some measures of variability is the range, variation ratio and standard deviation / variance.

Lab 2 Key Takeaways

1 | Reading the Data

In section 1, we read the data using pandas and then access information about the dataset and it's features using .info() and .describe()

2 | Representing the Data

In section 2, we look for the number of unique values, convert variables into categorical or numerical using .astype() if needed, and look for any places to perform data visualization.

3 | Missing Values

In section 3, we perform missing value imputations using the mean, median, mode and KNN. We create kernel density plots and determine which method of imputation would be the best.

Data Visualization

Data Visualization

Exploratory Data Analysis

We use EDA to search for patterns and trends in a given data set. The goal here is to suggest interesting hypothesis and find violations of statistical assumptions. We look for trends over the dataset that may not be obvious from the summary statistics.

1 Variable Visualization

For visualization of single variables, we can use histogram or density plots for numerical data or bar charts for categorical data.

2 Variable Visualization

To compare two variables of which both are numerical, we can use scatterplot, jointplot or a kernel density plot. To compare two variables of categorical and numerical, we can use boxplots. Else, to compare two variables of which both are categorical, we can use a heatmap.

Random Variables

Random Variables

Bernoulli & Binomial Random Variables

A Bernoulli Random Variable is used to model the outcome of an experiment that can be answered as a boolean. A Binomial Random Variable is used to model the sum of n Bernoulli experiments, such as how many successes after n tries?

Normal Random Variables

A Normal Random Variable can be used to model about any interval or ratio data. It has a probability distribution with two parameters, the mean and standard deviation.

Lab 3 Key Takeaways

1 | Reading the Data

In section 1, we need to identify the data types and and convert any variables into their respective data types. We also need to validate as to why those data types are the ones that were identified.

2 | Classification

In section 2, we need to identify the a response variable and create the different types of visualizations that are needed. We use those visualizations to find any predictive relationships. We also look at the method of imputations to and see if it affects the existence of a predictive relationship. We also discuss about which variables could be included for a machine learning model.

3 | Comparing Features

When we do machine learning, if two features are highlight correlated to each other, it is of limited value to include both as predictors.

Bootstrapping

Bootstrapping

Approach to Bootstrapping

Bootstrapping is about sampling from the data with replacement, and compute the statistic and repeat it over again. Confidence interval,

Confidence Intervals

To measure the variation of a statistic, we use a confidence interval. A 95% confidence interval is a region around the statistic.

3 Steps to Bootstrapping

There are three steps to bootstrapping that we need to follow:

- Sample from the original data with replacement and compute the statistic of interest 1000 times
- Sort the results from smallest to largest
- Take the 25th and 975th results from this sorted list - these are the boundaries of the confidence interval

Hypothesis Testing

Developing a Research Question

A good research has three characteristics, which are listed as follows:

- **Clear** so that it provides enough specifics so that one's audience can easily understand it's purpose
- **Focused** so that is it narrow enough that it can be answered thoroughly
- **Specific** so that it explicitly identifies variables of interest and their relationship

Developing a Hypothesis

A hypothesis is a "educated guess" that answers a research question. There are three parts:

- **Null Hypothesis:** usually a claim that there is no effect or nothing of interest]
- **Alternative Hypothesis:** usually a claim that there is an effect
- **Test Statistic:** a point estimate summary of the data allowing us to decide between null & hypothesis
- Rejection Criteria: if this happens to the test statistic, then we will decide we have falsified or rejected the null hypothesis

Hypothesis Testing

Type I Error & P Value

A type I error is when we reject the null hypothesis when it is true. This is what we call a false positive. We use a p-value which is defined as the probability of getting a more extreme value than the observed test statistic given that the null hypothesis is true.

Rejecting & Accepting Null Hypothesis

The usually choose the cut-off for the significance value as 0.05. We reject the null hypothesis when we are sufficiently unlikely to be making a type I error and is less than the cut-off of 0.05.

Lab 4 Key Takeaways

1 | Plotting Normal Distribution

In section 1, we instantiate a normal distribution and plot the probability density function (PDF), cumulative distribution function (CDF) and inverse cumulative distribution function.

2 | Method for Sampling

In section 2, we generate histograms for normal distributions of a number of samples. For each of these histograms, we directly plot the normal PDF over the histogram.

3 | Type I Errors

In section 3, we create a function that accepts two distribution objects and in turn it returns three lists , one of which consists of effect sizes.

4 | Type II Errors

In section 4, we create the same function and repeat the same experiment. With a adjusted number of iterations, we plot a histogram of the estimated effect sizes.

Lab 4 Key Takeaways

5 | Bootstrapping

In section 5, we implement a simple bootstrapping algorithm from scratch to estimate the confidence interval around the median of the data set.

Statistical Testing

Statistical Testing

Two Sample t-Test

A two sample t-test is a hypothesis test that compares a categorical variable with two values against a numerical variable. It answers the question of whether the means of two groups defined by the categorical variable differ.

- **Null Hypothesis:** the means of two groups are equal
- **Alternative Hypothesis:** the means of two groups aren't equal
- **Rejection Criteria:** p-value < 0.05

Kruskal-Wallis Test

A Kruskal-Wallis test is a hypothesis test that compares a categorical variable with multiple values vs a numerical variable. It answers the question of whether at-least one group defined by the categorical variable differs from at-least one other.

- **Null Hypothesis:** medians of all of the groups are equal
- **Alternative Hypothesis:** the medians of at least two groups are not equal
- **Rejection Criteria:** p-value < 0.05

Statistical Testing

Pearson's Correlation

The Pearson's Correlation Coefficient measures the strength of the linear relationship between two numerical variables. It answers the question of whether two variables move together (positive correlation) or opposite (negative correlation). It lies between -1 and +1.

- **Null Hypothesis:** variables are uncorrelated so that R equals 0
- **Alternative Hypothesis:** variables are correlated so that R doesn't equal 0
- **Rejection Criteria:** p-value < 0.05

Pearson's correlation measures the strength / direction of the linear relationship between two numerical variables. If Pearson's correlation is near 0, it might conclude that there is no relationship between the variables, but there could be a non-linear relationship.

Spearman's Correlation

The Spearman's Correlation Coefficient is a non-parametric measure of the monotonic relationship between two variables. It assesses whether an increase in one variable generally corresponds to an increase or decrease in another, even if the relationship is not linear.

- **Null Hypothesis:** there is no monotonic relationship between the two variables
- **Alternative Hypothesis:** there is a monotonic relationship between the two variables
- **Rejection Criteria:** p-value < 0.05

Statistical Testing

χ^2 Test of Independence

A χ^2 test of independence is a hypothesis test that compares two categorical variables and measures whether there is 'dependence' between them.

- **Null Hypothesis:** there is no association between the two categorical variables
- **Alternative Hypothesis:** there is an association between the two categorical variables
- **Rejection Criteria:** $p\text{-value} < 0.05$

Family Wise Error

Family Wise Error

Probability of Atleast One False Positive

If you do one test the probability of no false positive is 1 - α. If you do k tests the probability of no false positives is (1 - α)^k. So, if we do k tests the probability of at least one false positive is 1 - (1 - α)^k.

Statistical Families

We can count all tests in the same statistical family together. A family is:

- Multiple variables which are being tested with no pre-defined hypothesis
- Multiple tests together help support the same research question
- Could be tests conducted simultaneously or sequentially over a long period of time

For a family of k-tests 1 - (1 - α)^k is called the family-wise error rate!

Bonferroni Correction

Suppose we do k-tests simultaneously. We can reject the null hypothesis if the p-value is less than or equal to α / k. This guarantees that the probability of greater than or equal to type I error with k tests is no more than 0.05.

Sequential Correction Methods

α-Spending

In α-Spending, we set a wealth of W = 0.05. We require that the sum of the α's for all tests is less than or equal to 0.05.

α-Investing

In α-Investing, we set a initial wealth of W. We update the wealth using a a formula, where the wealth for hypothesis testing grows when we get significant results and decreases when we don't.

α-Debt

In α-Debt, we set an initial α of 0.05. For each new test, we apply a Bonferroni correction that treats the family as all previous tests.

Lab 5 Key Takeaways

1 | Reading & Preparing Data

In this section, we rename the columns in the reduced data set to names that are appropriate descriptors of the information contained in the codebook. We then plot a distribution of the variable and determine which method would be best to deal with those missing values.

2 | One Sample t-Test from Scratch

In this section, we compare a numerical variable against a fixed number. The goal is to assess whether the numerical variable is "different" from the number that has been specified.

3 | Two Sample t-test from Scratch

In this section, we compare a numerical variable against a categorical variable. The goal is to assess whether the numerical variable is "different" across the categories.

4 | Pearson's Correlation

In this section, we compare a numerical variable against another numerical variable. The goal is to assess whether the "two" variables "move" together in a significantly related way.

Lab 6 Key Takeaways

1 | Reading & Preparing Data

In this section, we read in the data and then prepare the data. We perform feature engineering by making five new variables.

2 | Life Expectancy

In this section, we explored which variables are predictive of US county level life expectancy. Hence, we used the life expectancy level as a dependent variable. We tested for association between categorical and numerical variables using a Kruskal-Wallis test.

3 | Classification of Above Average Life Expectancy

In this section, we explored which variables are predictive of US county level life expectancy. We ran a Kruskal-Wallis test for each numerical variable versus the `above_average_life-expectancy`. We can also test two categorical variables for association using a χ^2 test of independence.