DATA SCIENCE AND MACHINE LEARNING PIPELINE

A Beginner’s Guide to Architecting DS and ML Pipeline

Kennedy Kamande Wangari

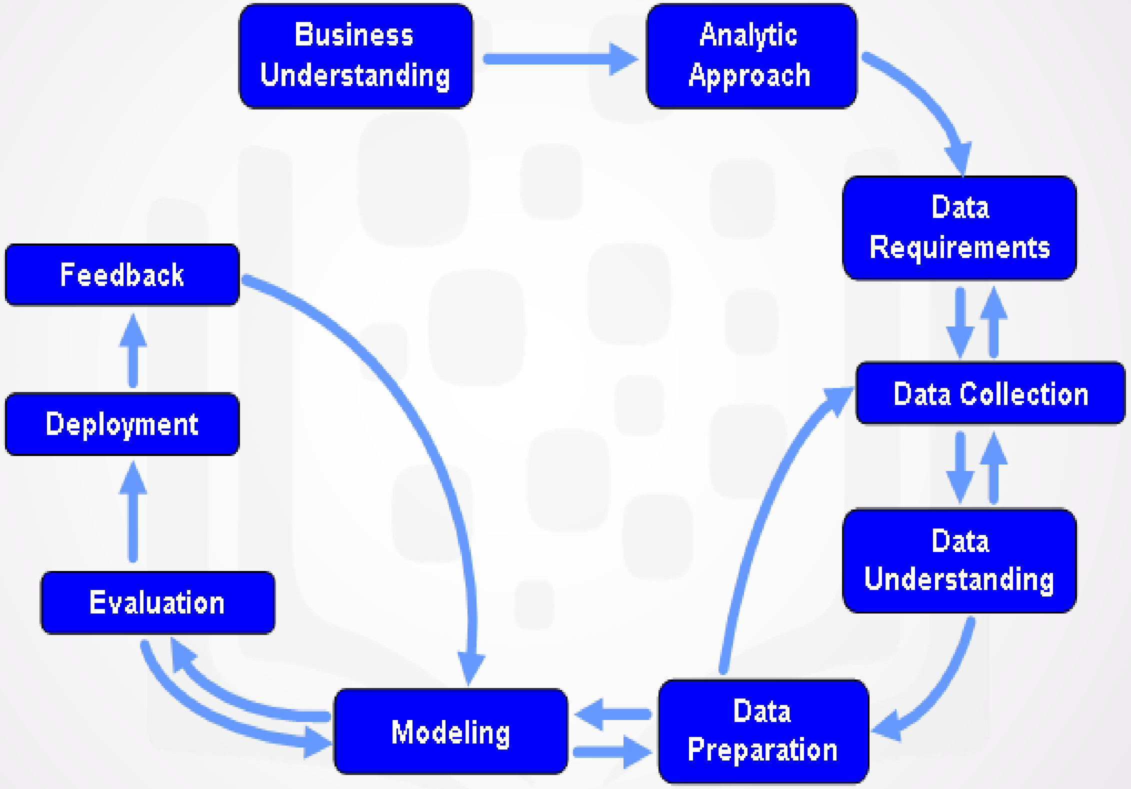
Data Scientist, Commercial Bank of Africa, A.I. Collaborator, Omdena and Active Kaggler.

https://www.linkedin.com/in/kennedykwangari/
https://www.kaggle.com/kennedywangari13432

Table of Contents

- Typical Data Science Work Flow
- Challenges in the Data Science Work Flow
- Exploratory Data Analysis
- Feature Engineering
- Model Building and Performance Evaluation
- Model Deployment to Production

CRISP-DM Methodology diagram



Cross Industry Standard Process for Data Mining

Business Understanding

- This is the first, most crucial and important step to solving any data science problem and involves formulating the questions that you will use the data to solve.
- Here we attempt to understand the problem we are trying to solve
- **ASK YOURSELF:**
- How can we translate data into dollars?
- What impact do I want to make with this data?
- What business value does our model bring on the table?
- What will save us lots of money?
- What can be done to make our business run more efficiently.

- Mastering this fundamental concept will lead you to greater steps in being successful towards your Data Science trajectory. No matter how well your model predicts, no matter how much data you acquire and no matter how great your exploratory data analysis is…….. **your solution or actionable insight will only be as good as the problem you set for yourself.**

- *Good data science is more about the questions you pose of the data rather than data munging and analysis." — Riley Newman*

- For example, you have gathered the data from online surveys, feedbacks from regular customers, historical purchase orders, historical complaints, past crises, etc. Now, using these piles of different data you may ask your data to answer the following:

- What should be the realistic sales goals for next quarter?
- What would be the optimal level of stock to have for the coming holiday season?
- What type of measures should the company take to retain customers?
- What can be done to minimize complaints?
- How can we bridge the gap between qualitative and quantitative matrices?
- What can be done to bring more happy customers?

- The more questions you ask of your data, the more insight you will get. This is how your own data yields hidden knowledge which has the potential to transform your business totally.

Hypothesis Generation

- This is the process of brainstorming and listing out all the possible factors that can affect the outcome.
- A hypothesis is a possible view or assertion of an analyst about the problem he/she working upon that maybe true or not true.
- This is done by understanding the problem statement in detail and thoroughly before embarking on the data.
- **For example, if you are asked to build a credit risk model to identify which customers are likely to lapse are which are not, these can a possible set of hypothesis:**
- Customers with poor credit history in past are more likely to default in future.
- Customers with high (loan value / income) are likely to default more than those with low ratio
- Customers doing impulsive shopping are more likely to be at a higher credit risk
- At this stage, you don't know which out of these hypothesis would be true.
- Conduct hypothesis driven data analysis.

Getting the system ready and Obtaining the data

- We will be using Python for our data science pipeline along with the below listed libraries.
- Pandas
- Numpy
- Scipy
- Matplotlib
- Seaborn
- ScikitLearn
- Data science can’t answer any question without data. First things first, obtain your authentic and reliable data.
- Identify all your available datasets(from the internet, external/ internal databases)
- Extract your data into a usable format (.csv, json, xml, etc)

- We will be using Python for our data science pipeline along with the above mentioned libraries.
- Download, install and configure Anaconda Distribution
- Loading the Python libraries.
- Reading the data from the data sets.

Understanding the Data

- In this section, we look at the structure of the dataset. Firstly, we will check the features present in our data and then we will look at their data types.
- We conduct Variable Identification to identify Predictor(Input) and Target(Output) variables.
- We also explore the shape our dataset and identify the category of the variables.
- Types of variable (Predictor or Target variable)
- Data types of the features(character, Numeric)
- Category of the variables(Continuous and Categorical)

Data Preparation

- This encompasses the activities to construct and clean the data set. This is because the results and output of your machine learning is only as good as what you input into it. Again garbage in, garbage out.
- Most of the times, data comes with its own anomalies like missing parameters, duplicate values, irrelevant features etc. So examine the data to understand every feature then do a cleanup exercise and take the information that is important to the problem asked.
- Clean the data: filling the data holes, remove duplicate or corrupt records, throwing away the whole feature sometimes.. Etc. Domain level expertise is crucial at this stage to understand the impact of any feature or value.
- Missing data reduces the power/ fit of the model or lead to a biased model because we haven't analyzed the behavior and relationship with other variables correctly, thus leading to wrong prediction or classification.
- Data wrangling tools: Pandas
- Deletion, Mean/ Mode Imputation and KNN Imputation methods are used to treat the missing values

- ***The man who is prepared has his battle half fought"—Miguel de Cervantes***

Outlier Detection and Treatment

- Outlier is a commonly used terminology by analysts and data scientists as it needs close attention else it can result in wildly wrong estimations. Simply speaking, Outlier is an observation that appears far away and diverges from an overall pattern in a sample.
- This is an observation that appears far away and diverges from an overall pattern in a sample.
- Let’s take an example, we do customer profiling and find out that the average annual income of customers is $0.8 million. But, there are two customers having annual income of $4 and $4.2 million. These two customers annual income is much higher than rest of the population. These two observations will be seen as Outliers.
- Outliers tend to make your data skewed and reduces accuracy.
- They increase the error variance and reduces the power of statistical tests.
- They bias and influence estimates that may be of substantive interest and impact the basic assumption of Regression, ANOVA and other statistical model assumptions.

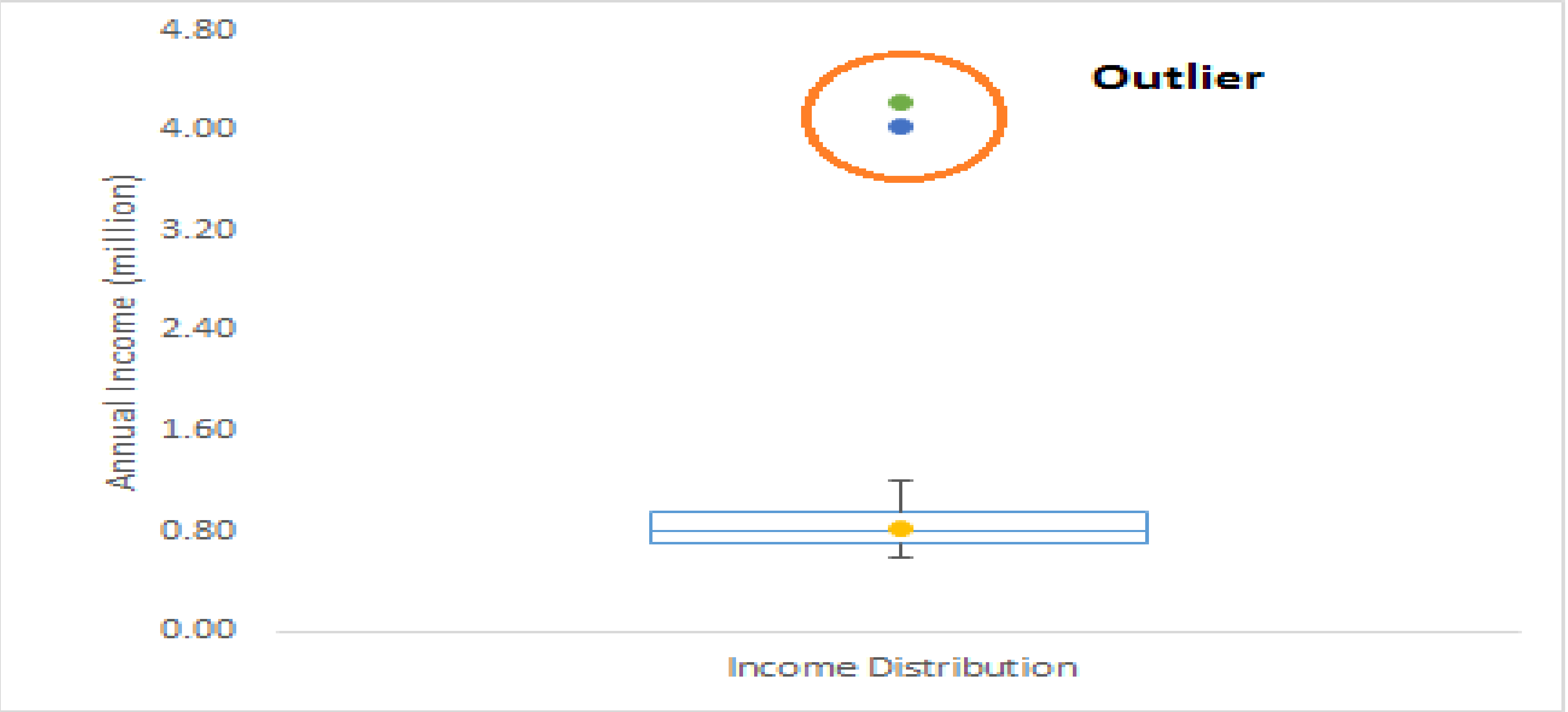
Exploratory Data Analysis (EDA)

Univariate Analysis

- At this stage, we explore variables one by one individually. This is the simplest form of analyzing data. The method to perform univariate analysis depends on whether the variable is categorical or continuous.
- Method to perform univariate analysis will depend on whether the variable type is categorical or continuous. For categorical features we can use frequency table or bar plots which will calculate the number of each category in a particular variable. For numerical features, probability density plots can be used to look at the distribution of the variable.
- Continuous variables: we need to understand the central tendency and spread of the variable. This is measured using various statistical metrics visualization methods. (Descriptive Statistics Chips in)
- Categorical variables: we'll use frequency table to understand distribution of each category. We can also read as a percentage of values under each category using two metrics: **Count** and **Count%.**
- For visualization we use bar charts

Bivariate Analysis

- Bi-variate Analysis finds out the relationship between two variables. We look for association and disassociation between variables at a predefined significance level. We perform bi-variate analysis for any combination of categorical and continuous variable. The combination can be: Categorical & Categorical, Categorical $ Continuous and Continuous & Continuous. Different methods are used to tackle these combinations during data analysis process.
- We test the hypothesis generated earlier, explore the variables with respect to the target variable.
- **Categorical Independent Variable vs Target Variable**
- Lets recall some of the hypotheses that we generated earlier:
- Applicants with high income should have more chances of loan approval.
- Applicants who have repaid their previous debts should have higher chances of loan approval.
- Loan approval should also depend on the loan amount. If the loan amount is less, chances of loan approval should be high.
- Lesser the amount to be paid monthly to repay the loan, higher the chances of loan approval.
- Lets try to test the above mentioned hypotheses using bivariate analysis.



We remove outliers through deleting the observations, transforming them, binning them, treat them as a separate group, imputing values.
These methods are similar to methods of treating the missing values

The Art of Feature Engineering

- This is the science (and art) of extracting and bringing out more information from existing data. You are not adding any new data here, but you are actually making the data you already have more useful.
- For example, let's say you are trying to predict foot fall in a shopping mall based on dates. If you try and use the dates directly, you may not be able to extract meaningful insights from the data. This is because the foot fall is less affected by the day of the month than it is by the day of the week. Now this information about day of week is implicit in your data. You need to bring it out to make your model better.
- **What is the process of Feature Engineering ?**
- Variable transformation.
- Variable / Feature creation.
- These two techniques are vital in data exploration and have a remarkable impact on the power of prediction.

What is Variable Transformation?

- In data modelling, transformation refers to the replacement of a variable by a function. For instance, replacing a variable x by the square / cube root or logarithm x is a transformation. In other words, transformation is a process that changes the distribution or relationship of a variable with others.
- Let’s look at the situations when variable transformation is useful.
- Variable transformation is useful when we want to change the scale of a variable, transforming complex non linear relationships into linear relationships and when symmetric distribution is preferred over skewed distribution.
- It is easier to interpret and generate inferences when using symmetric distributions of variables.
- For skewed distributions we can use transformations to reduce skewedness. For right skewed distribution, we take square / cube root or logarithm of variable and for left skewed, we take square / cube or exponential of variables.

Feature / Variable Creation and Its Benefits

- This is the process of generating new variables / features based on existing variable(s).
- For example, say, we have date(dd-mm-yy) as an input variable in a data set. We can generate new variables like day, month, year, week, weekday that may have better relationship with target variable.
- This step is used to highlight the hidden relationship in a variable:

Techniques used to create new features are

- **Creating derived variables:** This refers to creating new variables from existing variable(s) using set of functions or different methods. Let's look at it through "**Titanic – Kaggle competition**". In this data set, variable age has missing values. To predict missing values, we used the salutation (Master, Mr, Miss, Mrs) of name as a new variable. How do we decide which variable to create? Honestly, this depends on business understanding of the analyst, his curiosity and the set of hypothesis he might have about the problem. Methods such as taking log of variables, binning variables and other methods of variable transformation can also be used to create new variables.
- **Creating dummy variables:** The most common application of dummy variable is to convert categorical variable into numerical variables.
- Dummy variables are also called Indicator Variables. It is useful to take categorical variable as a predictor in statistical models. Categorical variable can take values 0 and 1. Let's take a variable 'gender'. We can produce two variables, namely, "**Var_Male**" with values 1 (Male) and 0 (No male) and "**Var_Female**" with values 1 (Female) and 0 (No Female). We can also create dummy variables for more than two classes of a categorical variables with n or n-1 dummy variables.

Automated Feature Engineering

- Feature Engineering is very crucial when it comes to machine learning hackathons and competitions. Its often the difference between getting into the top 10 of the leaderboard and finishing outside below the top 50!.
- The performance of a predictive model is heavily dependent on the quality of the features in the dataset used to train that model. If you are able to create new features which help in providing more information to the model about the target variable, it's performance will go up. Hence, when we don't have enough quality features in our dataset, we have to lean on feature engineering.
- It has immense potential, but it can be sow ad arduous process when done manually. You have to spend time brainstorming over what features to come up and analyze their usability from different angles. You can automate the entire process.
- Automating feature engineering makes the machine learning building process much more efficient and cost effective, thus allowing the data scientist to focus on other aspects of the model.
- Python offers a great tool to address automated feature engineering through a library called Featuretools.

Model Building Process

- This is the juiciest and funniest part of the data science pipeline.
- After cleaning your data and finding what features are most important, using your model as a predictive tool will only enhance your business **decision making**.
- The predictive model is created via a process called “training”. The goal of training is to create an accurate model that answers our questions correctly most of the time.
- Rule of the thumb: split your clean data set in the order of 80/20 or 70/30 into two parts. The 80/70 % data set is used for training the model and the 20/30% data set as test data for evaluating the performance of the model.

- **Choosing a model**
- The next step in our workflow is choosing a model. There are many models that researchers and data scientists have created over the years. Some are very well suited for image data, others for sequences (like text, or music), some for numerical data, others for text-based data.
- Supervised, Unsupervised and Reinforcement Learning Machine Learning Algorithms
- **Training**
- This is the bulk of machine learning. We use the test data to incrementally improve our model's ability to predict.
- **Evaluation**
- Once we have completed the training process here we use the test data set that we had set aside earlier to test our model against. This metric allows us to see how the model might perform and is meant to be a representative of how the model will perform in the real world.
- **Hypermeter Tuning**
- Once done with the evaluation process, we hypermeter tune to see if we can further improve the training of our model in any way.
- Here we tune the parameters. Here we consider the few parameters we had implicitly assumed when we did our training, thus we now go back and test those assumptions and try other values.

- **Prediction**
- Machine learning is using data to answer questions. So **Prediction**, or inference, is the step where we get to answer some questions. This is the point of all this work, where the value of machine learning is realized.

- **TensorFlow Playground**
- To learn ways to play with training and parameters, check out the TensorFlow Playground. It's a completely browser-based machine learning sandbox where you can try different parameters and run training against mock datasets.

What's next?

- While we will encounter more steps and nuances in the future, this serves as a good foundational framework to help think through the problem, giving us a common language to talk about each step, and go deeper in the future.