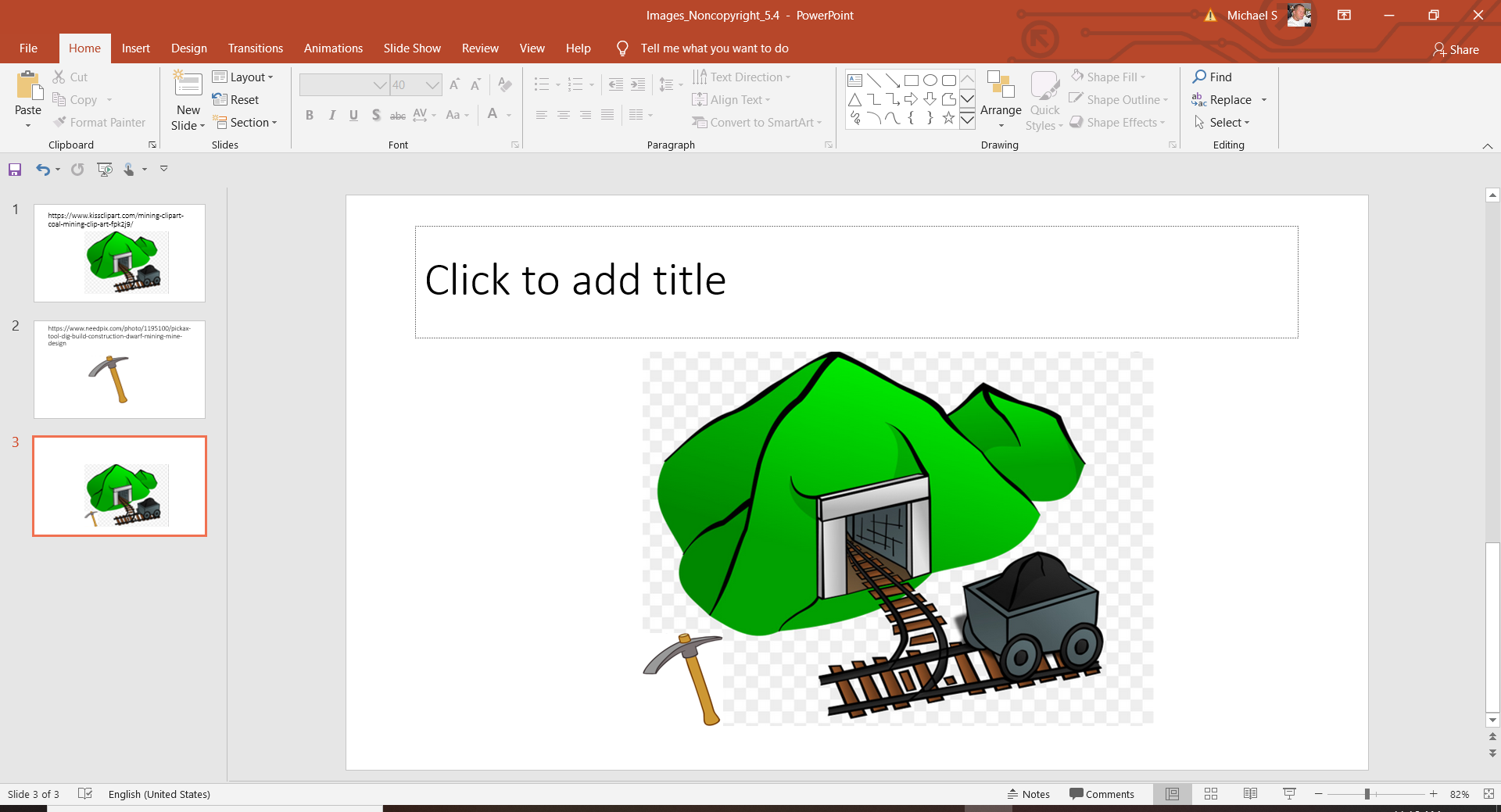
**Introduction to Data Mining**

Learning Goals

* **Define and discuss data mining and concepts of data mining**
* **Discuss common data mining techniques to gain insight and reveal trends and patterns**
* **Discuss cross-industry standard processes for data mining**
* **Gain understanding of what industries use data mining and for what purposes**
* **Conduct data mining - cluster analysis**

Introduction

The term “data mining” can inspire confusion if you are just starting out in your quest to become a data analyst. For example, when you think of gold mining, you might imagine a large operation with people using pickaxes and machines to dig up the dirt, then apply various techniques to identify the gold and separate it from the rest of the dirt. The objective of gold mining is to find gold, of course! But with data mining the objective is not to find data. As you remember from Introduction to Big Data, there is already plenty of data. The objective of data mining is to discover trends, patterns and insights from data that in turn allows the users to make more informed decisions. Much like a miner applies tools, such as a pickaxe, to the dirt to discover gold, a data analyst applies quantitative methods to data to discover patterns and insights. In this achievement, we will further explore the concept of data mining, review some techniques of data mining, standard data mining methodologies, as well as discuss how knowledge discovered through data mining is applied.

Methodology (CRISP-DM)

When it comes to data mining and data analytics for that matter, there are very few absolutes. Meaning, there is no absolute right way to conduct data mining. If you were to talk to data analysts working in different places, you would find that their methodologies of data mining can vary from company to company and industry to industry. Some methods may have more steps than others, or combine or separate certain tasks into different stages, but the goal is the same. However, to ensure consistency, your team should adopt a standard methodology to data mining. The Cross-Industry Process for Data Mining (CRISP\_DM) is a thorough and validated approach to data mining. It is a structured approach to data mining and breaks the process down into the following six steps: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment.

Business Understanding

The Business Understanding phase simply means to understand the specific goals of the business, your department’s role in the business, and tailor your data mining projects to support them. If your work is to help identify and assess risk, it should be easily understood how your project will do that. If your work supports the business goal of cutting costs through increasing efficiencies, your projects should answer exactly how it will result in cost savings.

You may often find yourself waiting on other colleagues who have competing priorities and other obstacles you will need to work around in a typical work setting that is going to extend the time you need to complete your project. Understanding your business goals and priorities will help you avoid using valuable time leading projects that produce answers to questions nobody is asking. Be able to define success for your projects. Success isn’t just completing the project. Depending on the nature of your work, your expected results should be measurable and one should be able to quantify the impact to your business’s goals.

Data Understanding

At this phase you will collect your data and make some initial analysis of your data. Begin to document everything about your project and data. By the time your project is complete, a colleague should be able to peer review your work and know there were no missed steps or flaws in the tests. Document the methods in which the data was collected, data quality issues, and potential data bias. Execute some queries against the data and validate the data is consistent with your query or data request. Create some visuals that reflect the quality, quantity, and notable characteristics of the data.

Data Preparation

The Data Preparation phase includes the activities you execute to create a dataset ready to be analyzed from your data. Select and clean your data and document the data quality and integrity. If quality or integrity are issues, include the steps taken to alleviate them. Based on your project plan you may need to construct derived attributes to add to your table. For example, your table contains columns for number of transactions and aggregate amount of transactions, but no column for average amount per transaction. Construct the data you require using the appropriate tools and formulas.

Modeling

The goal and nature of your project will determine what combination of logic and algorithm or algorithms to be applied to your dataset. Some of the more common algorithms are decision-tree, clustering/segmentation, regression, sequence and classification. Be sure you are documenting which techniques are used and any assumptions they make. Create a procedure or process for testing your model, also called independent model validation. As a data analyst working for a global financial services company, your job is to develop models that monitor transaction activity and predict money laundering risk. The output generated by your model will need to be tested in a structured and systematic way to ensure the output is consistent with the logic, parameters as well as assess the rate of accuracy. With your algorithms selected and test designed, construct your prototype models.

Evaluation

By this time, you have constructed your prototype models and produced output that was validated by a team of subject matter experts. Their aggregate scores will determine if any model or models will be recommended for implementation, further testing, or abandoned. Create visuals of your results so they can be further analyzed. Take any other feedback from your validation team that helps you understand if the model is sound or if some tinkering with logic and parameters is warranted prior to re-testing. Review and document your mining process and prepare visuals and presentation materials to support your recommendations.

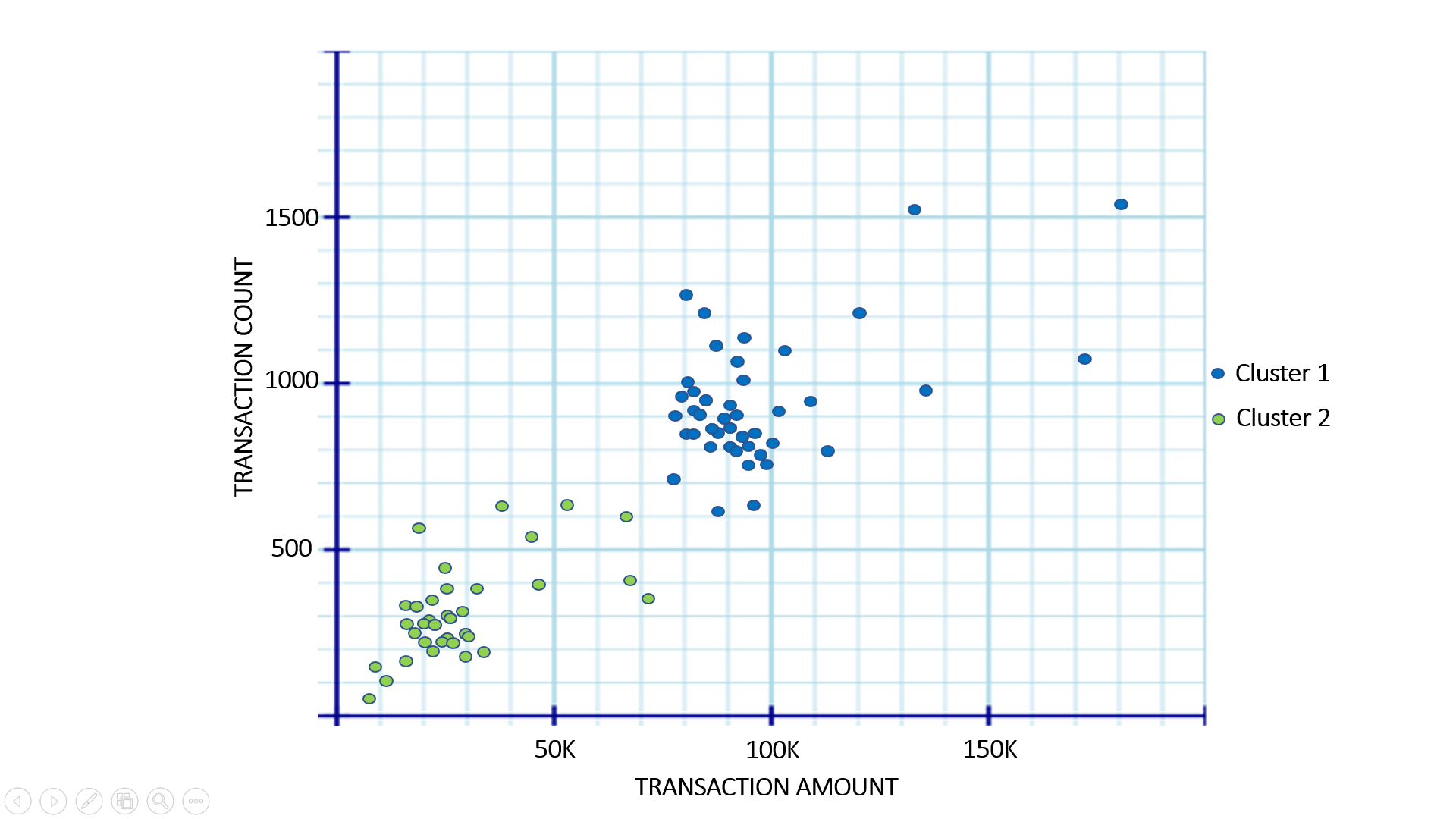
Deployment

Congratulations! One of your anti-money laundering prototype models produced an accuracy rate for identifying risky transactions that is good enough to make your Compliance Committee decide to implement it into the bank’s transaction monitoring program. Your model indicates it will deliver results consistent with your business needs, which is to identify and report suspicious transactions more efficiently and cost effective. Your model will be generating output on a daily basis that will be investigated and validated. That validation is input into your model so that your model’s performance can be tracked. Continue to monitor and document model performance and investigate any radical changes to output volume or quality. Generate periodic reports as required by your business that detail model development, purpose, quantitative threshold changes and performance.

Common Data Mining Algorithms

Clustering

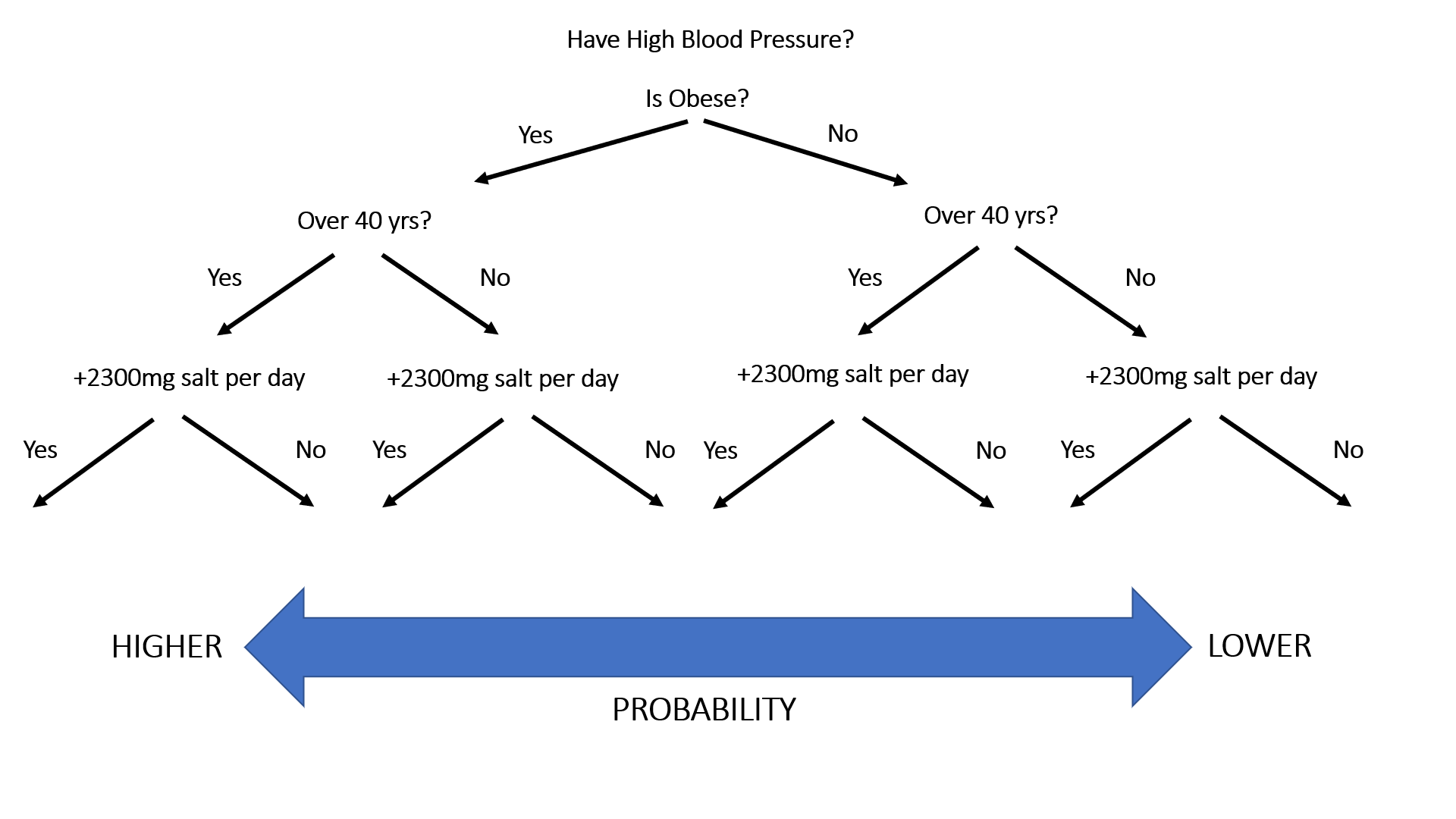
You work as a data analyst working in the AML Compliance Department of a bank. Your bank currently uses many different models that monitor and alert for transaction patterns that may indicate laundering. The problem is that the models all apply the same monetary thresholds to each client’s transaction activity, which leads to a higher false positive rate of your model outputs. Knowing that it’s natural for some clients to transact more and for higher amounts than others based on various factors, your bank has decided this “one size fit all” approach is inefficient. Your bank wants your team to find a way to divide the bank’s clientele into different groups so that the appropriate transaction thresholds can be applied based on the level of transaction activity expected of each client segment. It is reasonable to assume the accounts of a 45 year old high school teacher and large retail business are going to be quite different in terms of transaction volume and amounts. After discussion with your team, you have decided to start run a clustering algorithm with 2 centroids against the transaction data of the bank’s clients to support the segmentation project.

The clustering algorithm is commonly used by data analysts in many industries to identify like-subgroups (clusters) within the data. Clustering is applicable to client or market segmentation, or can be used for more advanced data exploration. One of most useful clustering algorithms is the Kmeans. It works to partition your data into distinct and predefined number of subgroups by reducing the sum of the squared distance between the means and data points. Then each data point is assigned to the closest mean, and in affect forming a cluster around the means. Your visual tools can be used to better understand your clusters. 

Your visual built on your clustering algorithm has revealed two clear clusters that reveal some deep differences in the transaction frequency and amounts of the bank’s clientele. After further analyzing your dataset, you find the majority of the majority of green dots are accounts owned by individuals and the blue dots are mostly business accounts. Armed with this knowledge, you can now recommend segmenting your bank’s clientele into individual and business so that the more optimal model thresholds can be applied to the appropriate segment. From here, you can run more clustering algorithms to segment your individuals and business clients to more granular levels.

Decision Tree

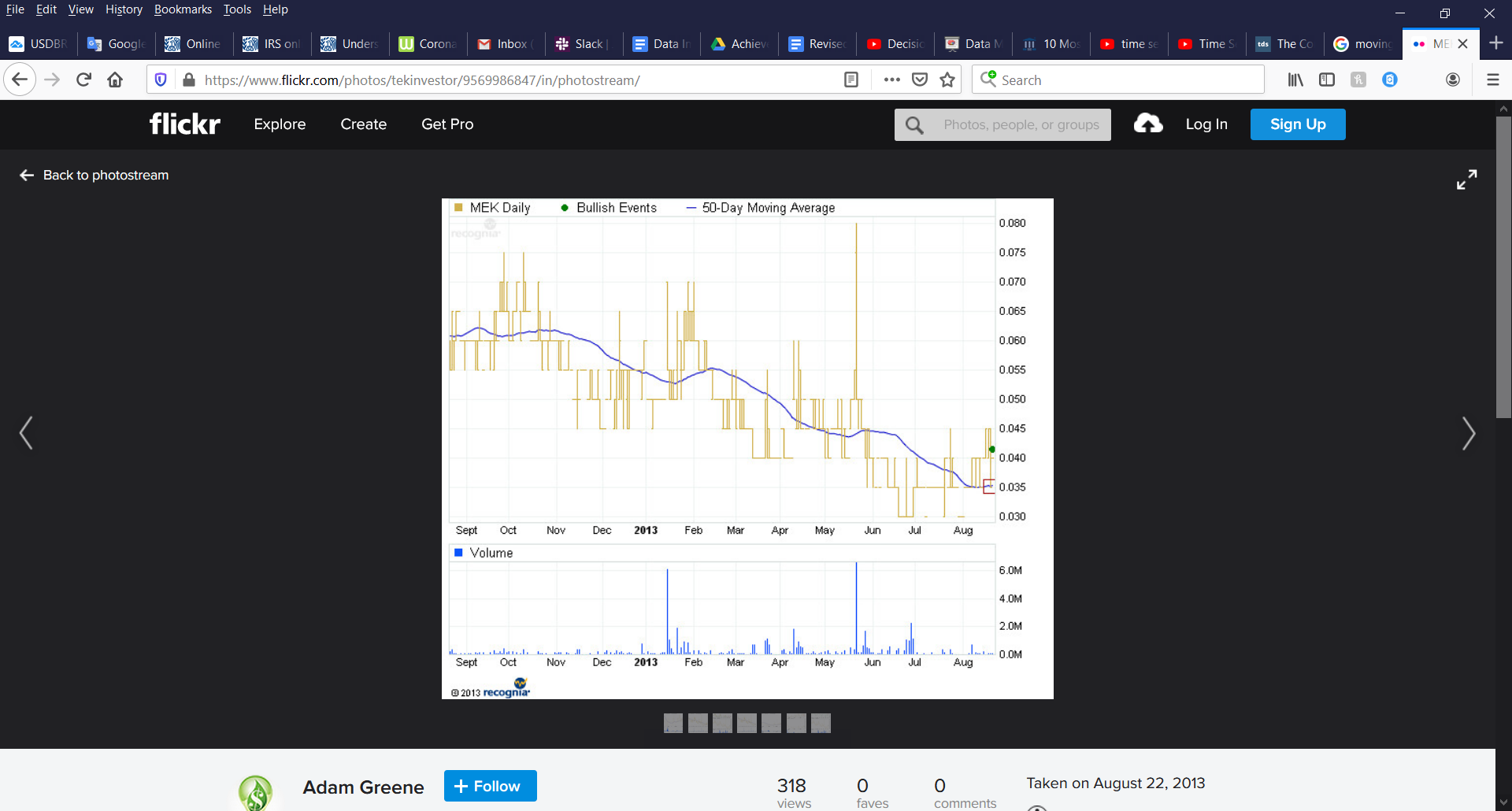
A decision tree is an algorithm used to support decision making, classifications and estimations. The decision tree models different outcomes, probabilities and consequences by breaking down data points into smaller subsets as it “passes through” through a linear series of questions. Each decision node branches into two or more different outcomes, giving the decision tree visual a tree-like appearance. The nodes towards the top will have the greatest impact on outcome.



Imagine you are a data analyst working for a medical facility and you are using a table including patients and their medical information. You are tasked to build a model that estimates the probability a new patient will have high blood pressure based on the variables you include in your model. The above image is a basic example of a decision tree you might use to determine the probability based on the affirmative or negative responses to the decision nodes. Decision trees are relatively easy to construct and understand

Time Series

The time series is an algorithm often used for forecasting. It is equal to a regression model in that it seeks to predict quantitative values based on previous values. One way time series can be modeled is by the moving average. A moving average simply attempts to predict the next value in a series of values by averaging the previous values over a given amount of time. The visual effect of the moving average is to smooth out larger movements in value to discover broader trends. It is commonly used by financial analysts and traders to identify market trends, assess volatility and help determine when assets are under or overvalued.



The yellow line in the image above represents movements in MEK price action over 12 months. The blue line is a 50-day moving average. In other words, the average closing price of the previous 50 days. This gives the analyst a better insight into the price movement without the noise of day to day changes.

Conclusion

Data mining is the meat and potatoes of data analytics. Taking raw data and applying quantitative methods to discover trends and actionable information is fundamental to being a data analyst. We have explored concepts and process of data mining, some commonly used algorithms, and some real-world application of each. It is important to remember that data mining has nothing to do with searching for data. Data mining is about discovering insight into the world around us by applying mathematical concepts to data. When you work within trusted processes such as CRISP-DM, and ensure your projects are aligned with your industry’s goals, you are going to maximize the effectiveness of your data mining.

Resources

<https://www.sv-europe.com/crisp-dm-methodology/#dataunderstanding>

<https://www.datasciencecentral.com/profiles/blogs/crisp-dm-a-standard-methodology-to-ensure-a-good-outcome>

<https://www.datasciencecentral.com/profiles/blogs/crisp-dm-a-standard-methodology-to-ensure-a-good-outcome>

<https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a>

<https://medium.com/@chiragsehra42/decision-trees-explained-easily-28f23241248>

<https://towardsdatascience.com/the-complete-guide-to-time-series-analysis-and-forecasting-70d476bfe775>

Images

<https://www.kissclipart.com/mining-clipart-coal-mining-clip-art-fpk2j9/>

<https://www.needpix.com/photo/1195100/pickax-tool-dig-build-construction-dwarf-mining-mine-design>