**Predictive Analysis**

Learning Goals

* **Define predictive analytics and explain how it's used (can draw comparison with descriptive analysis)**
* **Discuss the intersection between data analytics and data science**
* **Explain the business application of Predictive Analysis**
* **Discuss the role of classical statistics in modeling**
* **Know the steps of predictive analysis**

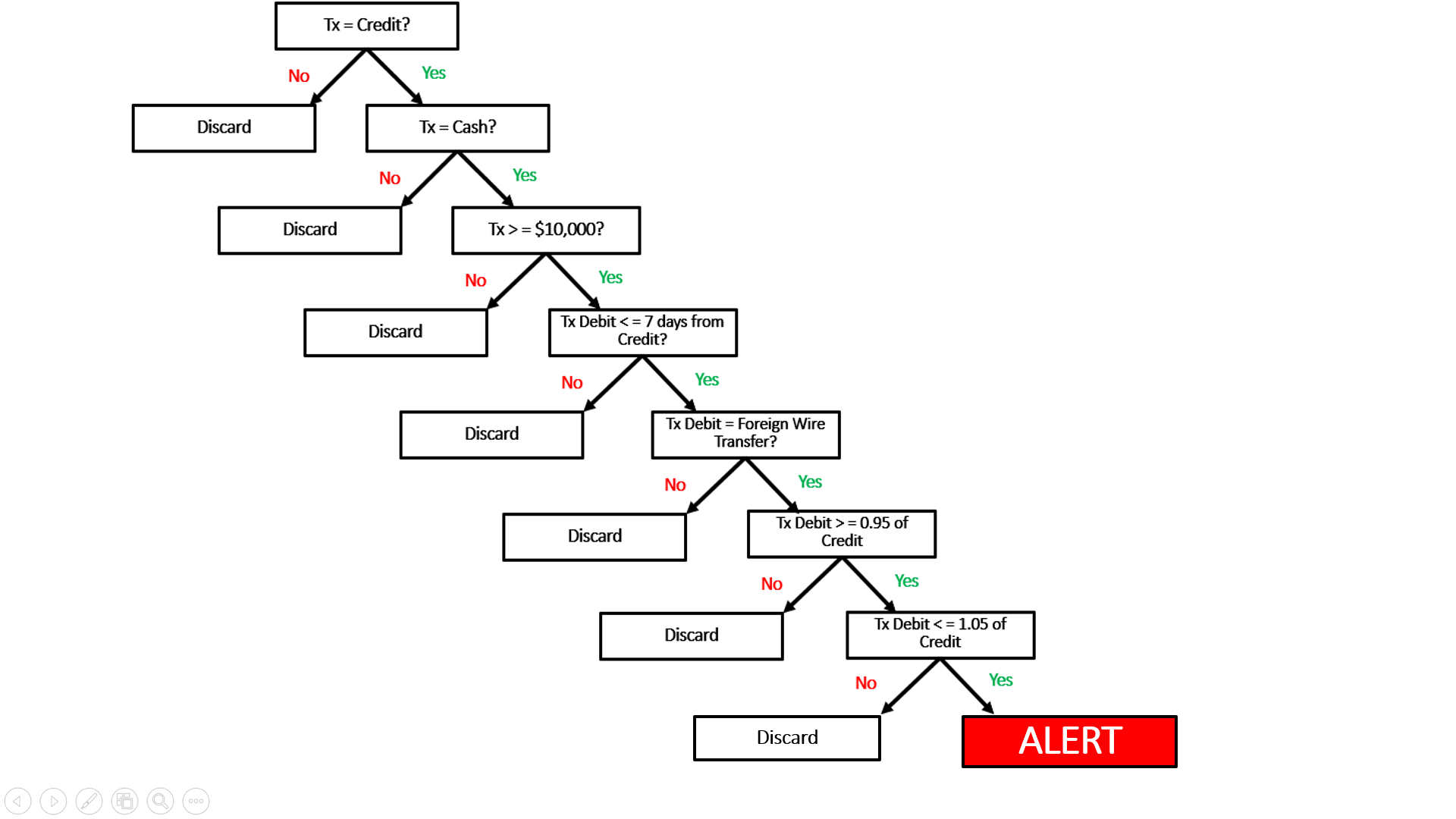
Introduction

In the ultra-ambitious world of business where profit margins are tight and competition is fierce, the need to stay ahead of the pack is greater than ever. Businesses are looking to data analysts to develop models to make predictions from which they can formulate strategies. Predictive analysis is the skill-set the data analyst uses to address these business needs. In section 5.4, we defined data mining as the process of applying mathematical algorithms with the objective of discover trends, patterns and insights. Predictive analysis can be defined as the application of those trends, patterns and insights found in historical data to build models which better predict outcomes with some degree of accuracy. It is fundamental to remain cognizant that a predictive model is built on the assumptions of statistics and historical data input into the model. Garbage in, garbage out as they say. Accurate predictive analysis is the result of many steps. In keeping with the idea of best practices, a standard process to predictive analysis should be followed. This is a key difference separating data analytics from data science. As a data analyst, you will be expected to adhere to proven processes and techniques for discovering insights and knowledge within data, whereas a data scientist is focused on testing and innovating new processes and techniques for data analysts to follow. In this section, we will discuss the concepts of predictive analysis, how to perform predictive analysis and applications of predictive analysis. Predictive analysis has almost limitless applications no matter what industry in which you may work.

Steps of Predictive Analysis

Predictive analysis is a process. Following a standard, approved procedure is the best way to get results that are both consistent and trustworthy. Many data analysts work in environments where their work will be subjected to peer review and/or questioned by internal and external auditors. Stakeholders will always want to know the intimate details of your project and you should be prepared to explain the processes you used to achieve your results.

1. **Define Project Scope and Goal**. The first step of a predictive analysis project will not be unlike the first steps of many process models. Define what it is that you and your team are going to do, why they are going to do it and how they’re going to do it. This is where you layout your plan. Meet with your team and communicate your intent of the project. Make your hypothesis and design the project to validate or refute your hypothesis. Ensure your projects objective is aligned with your business’s objectives, articulate expectations and define a successful outcome.
2. **Collect and Prepare Data**. In this step you will gather the data and build your dataset. Define the sources of data that will be used and know how the data was collected so you can be aware of any potential bias. Your data may come from multiple sources, so it is here where you will need to perform joins and clean your data. This part of the process is tedious and monotonous, but any model is only as good as the data on which it is built. Conclude this step by documenting your sources of data, data quality, and any other noteworthy observations of your data.
3. **Testing and Analysis.** Now that your data has been sanitized and organized into a finished dataset, it is ready to be tested and/or analyzed. The analysis will consist of using various statistical techniques from which to recognize patterns, occurrences, and validate your hypothesis. This phase may also have you review the results of tests you designed to validate your assumptions. For example, let’s say you work as a data analyst for a global financial services company developing models that predict if transaction patterns could be money laundering. You theorized that a particular transaction pattern involving large cash deposits followed by wire transfers to foreign countries would be a strong foundation from which to build a new model. You collected thousands of historical instances of this transactional pattern and distributed random samples of that output to subject matter experts. Your test consists of subject matter experts reviewing the output and indicating “positive” for suspicious activity, or “negative” for not. For the transaction examples dispositioned “positive,” you may use statistical methods to measure the relationship of certain variables to each other, such as the amount of the transactions or the time in between the cash deposit and the wire debit. After your analysis has determined the transaction variables that are most associated with suspicion, you know will have a better idea what logic, filters, and thresholds will go into your prototype model.
4. **Modeling**. Your testing and analysis have indicated the basic logic of your model would be effective and predicting money laundering risk. Now is the time to build and test several prototype models using historical data. This time your subject matter experts will be reviewing and dispositioning the output from your prototype model, not the output of your SQL query. As you get results back, you will use classical statistics to make inferences about relations between variables and how they affect outcome. Then you will either refine the prototype models or scrap them if they aren’t showing potential. It may take multiple model iterations to arrive at a final decision. Document the performance of all prototype models and prepare your chosen model for presentation to stakeholders. Decision tree models for predicting risk of a financial crime are common throughout the financial industry. But other predictive models commonly used include linear and logistic regressions, random forests, and neural networks. The image below represents a decision tree example of what your model might look like. As you read it from top to bottom, note the logical progression the model uses to filter out the specific transaction patterns being tested for money laundering risk. The $10,000 decision node is your amount threshold of the initial cash deposit. Essentially, this model is designed to predict with an acceptable degree of accuracy that a trained financial crimes investigator would find the specific transaction activity suspicious.



**Tx = Transaction**

1. **Deployment.** With the model tested and approved, it is ready to be deployed in a live environment. This is more of an engineering task, however once deployed it should be monitored closely to make sure it has been implemented properly and performing as expected. The smallest coding error in how the model is written can result in output that is not consistent with the desired output.
2. **Continuous Monitoring and Validation**. This is the final stage of modeling and where you will collect data on your model’s performance and assess efficiency. Model validation is an ongoing process to verify the model is performing as expected and is consistent with its objectives. Drastic changes in output volume from month to month may indicate the model isn’t functioning correctly. In this case you have a model built around monitoring a specific money laundering typology. Typologies are always evolving and as launderers change their methods, old models can become ineffective.

As more data is collected on the model performance, optimization activities can improve the efficiency of the model. Optimization is the process of discovering the threshold values that maximize model efficiency. For example, in our model above, the cash deposit represents the amount threshold parameter within the model. The model has a 5% True Positive rate in the first year after implementation. A more granular look at the data show 95% of your True Positives produced were cases in which the cash deposit was $15,000 or higher. Adjusting the threshold to $15,000 would then be the more appropriate threshold to maximize efficiency. All parameter thresholds should be optimized periodically as part of normal validation.

Applications of Predictive Analytics

Today businesses and scientists are using predictive analytics in an almost infinite number of ways. The evolution of predictive analytics and its business application has created strong demand for those with the skills. Much success in the business world is heavily dependent on effectively and consistently anticipating changes in market conditions, supply/demand and customer behavior. Predictive analytics is used heavily in the science fields for things such as modeling climate changes and anticipating the pace of the spread of diseases. Below are some examples of how predictive analytics is used in different business functions and industries.

Financial Modeling

Businesses, especially banks and investment firms, use financial models to help them make a variety of business decisions. These models are used to predict return on investments and calculate risks versus rewards of investing in or buying certain companies. Credit models provide a consistent method of deciding whether or not to lend money to an individual or business and how much they should be willing to risk. If you have ever applied for a mortgage, you will have noticed the bank will tell you the maximum amount you are qualified to borrow to buy that house. It’s a model that produces that amount and uses your credit history, current income and other variables from which to determine it. Models also help financial analysts forecast earnings and evaluate securities so they can know what to recommend to their clients.

Marketing

In order to sell a product, it goes without saying you have to invest in advertising. Companies practice segmenting their customer base into subgroups based on attributes such as gender, age, and purchasing habits. The data available on each of these customer groups allows them to better predict what advertising methods for a given product would be most effective for each subgroup. This allows them to tailor their ads to the target customers and maximize the impact of what they spend on advertising.

Forecasting

Have you ever heard or read in financial related news how economists predict the gross domestic product (GDP) to grow or contract x% next quarter? This isn’t just guesswork; it is predictive analysis. Economists use economic data from a variety of sources to identify trends such as past growth, unemployment trends, interest rates and tax rates. This analysis gives government policy makers and business leaders the insights they need for strategic planning.

Health Care

Predictive analysis is used extensively throughout the health care industry. Scientific researchers can better determine a person’s risk of developing a certain diseases or ailments based on lifestyle factors, genetics, and environmental conditions. This leads to better decisions at the clinical level and doctors making more informed recommendations for their patients.

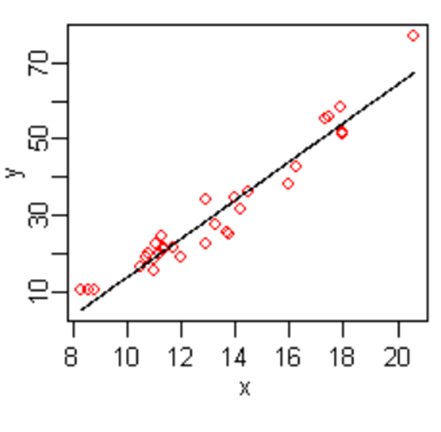
Regression for Predictive Analysis

Regression analysis at its core, is a statistical technique to measure the relationship between variables, usually for the purpose of creating a predictive model.  In simplest terms, it’s a way of explaining how one variable may affect another so that better predictions can be made.  Predictive analysis based on regression has many applications in numerous industries.  Scientists have used regression to measure the relationship between carbon dioxide in the atmosphere and temperature rise so they can develop predictive models.  Financial analysts use regression to measure the price relationships between different assets so they can determine the best prices to buy or sell them.  Operational managers need to know how production demands affect labor requirements and regression can tell them.  Regression is a versatile and proven technique to base your predictive analysis.

Methodology

One of the more common regression techniques to measure the relationship between a dependent variable and an independent variable is the linear regression.  Depending on which tools you are using to build a predictive model from your dataset, the step by step process can vary.  But the overarching process is the same and can be broken down into three steps.

1. **Make a hypothesis**.  Define what it is you want to investigate or test and make a reasoned assumption of what you believe will be the results.  For example, you hypothesize that the higher the average daytime temperature, the higher ice cream sales will be.  So your independent variable is going to be daily high temperatures, and your dependent variable will be daily ice cream sales at your local shop.  Next, collect price data from your shop and local temperature data and create your dataset.



1. **Create a scatter plot**.  With your dataset organized and cleaned, use your chosen software tool to create a scatter plot visual.  Ensure you label your x and y axis with the appropriate variable names.  Using your software tool’s functions, draw your regression line.
2. **Interpret the measurement and assess model fitness**.  Recall from your statistics classes that a slope can describe the relationship as strong/positive, weak/positive, strong/negative, weak/negative.  This will describe what relationship, if any, temperature may have on ice cream sales.  The r-squared will be a number between 0 and 1 and tells you how close the data are to your regression line.   The closer to 0 the r-squared, the less accurate your predictive model would be.  The closer the r-squared to 1, the better your model will be at predicting ice cream sales based on temperatures.  Looking at the image above, the close proximity of the datapoints to the regression line indicate a high r-squared value and better predictive model.

Conclusion

Predictive analytics is the culmination of a data analyst’s work.  It requires you to integrate your coding skills, understanding of data, command of statistics and data visualization skills to make accurate predictions.  Here, we have discussed methodology and concepts of predictive analytics, as well as the applications of predictive analytics in different industries.  Predictive analytics are often team efforts.  If you are leading the project be sure you are communicating your intent clearly and documenting your project milestones and results.  With your skills, predictive analytics methods, and leadership skills, both your career and your organization will benefit from predictive analytics.

Resources

<https://dzone.com/articles/introduction-to-predictive-analytics-and-predictiv>

<https://www.raconteur.net/technology/seven-stages-of-predictive-analytics-implementation>

<https://www.neuraldesigner.com/blog/6_Applications_of_predictive_analytics_in_business_intelligence#CustomerTargeting>

<https://medium.com/fintechexplained/part-3-regression-analysis-bcfe15a12866>

Images

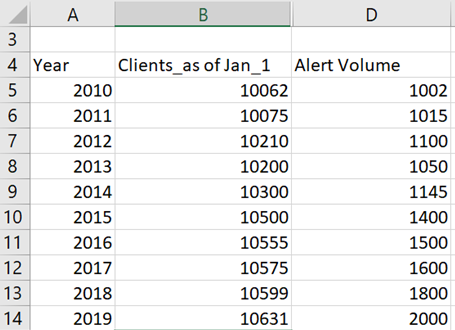
<https://medium.com/fintechexplained/part-3-regression-analysis-bcfe15a12866>

Task

**Background**:  As a data analyst for the Pig E. Bank Anti-Money Laundering Compliance Department, your job is to develop models that support monitoring transaction activity as well as provide analytics support to the operational processes of the department.  Your senior managers need to make personnel decisions based on how many alerts, or output, all of the models are creating so they can ensure the investigations section is adequately staffed to handle the workload.  You have been tasked with providing senior management a predictive model that accurately and consistently predicts the alert volume over time so that they can know when to hire more investigators.

**Process**

1. Hypothesis.  Your theorize that the annual number of alerts produced has a close relationship to how many clients the bank has.  Naturally, the more clients the bank has, the more transactions must be monitored and consequently more alerts will be produced.  You create the below dataset with the intention of doing a linear regression to measure the relationship between number of clients of the bank and the number of alerts produced over the past 9 years.



1. Create a scatter plot with your dataset then use the functions of your software tool to draw your regression line.
2. Complete your linear regression by showing equation and r-squared.

**Action**

You have now measured the relationship of number of clients to number of alerts.  Referencing the slope and r-squared of the above regression model, describe the relationship between the two variables.  Based on the results, would you advocate using this model to support predictive analysis?  Why or why not?