Lecture Note 1: Introduction to Business Analytics- Business Oriented Data Science

1. Business Analytics

1.1 My Definition

Business analytics is the process of collecting, analyzing, and interpreting data using technologies and data models to generate meaningful insights that facilitate better and faster decision making in business management.

Business analytics is a critical discipline in the modern business landscape, empowering organizations to harness the power of data and technology for informed decision-making and strategic planning. It encompasses a systematic and iterative process that involves the collection, analysis, and interpretation of vast and diverse datasets, transforming raw information into valuable insights and actionable knowledge.

At its core, business analytics revolves around the utilization of cutting-edge technologies, sophisticated data models, and statistical techniques to extract valuable information from structured and unstructured data sources. By leveraging data analytics tools and software, businesses can process and analyze vast amounts of information, identifying trends, patterns, and correlations that might otherwise remain hidden. These insights serve as a powerful resource for business leaders, enabling them to make data-driven decisions that can optimize processes, enhance performance, and capitalize on emerging opportunities.

The primary objective of business analytics is to facilitate better and faster decision-making in business management. By integrating data-driven methodologies into key aspects of organizational functions, businesses can effectively align their strategies with market demands, customer preferences, and operational needs. The ability to make informed decisions in a timely manner can be a significant competitive advantage, allowing companies to respond proactively to changes in the marketplace and gain an edge over competitors.

Furthermore, business analytics fosters a culture of evidence-based decision-making, mitigating the reliance on intuition and gut-feelings. Data-driven insights provide an objective foundation for discussions and strategic planning, increasing the likelihood of successful outcomes and minimizing the risks associated with uncertainty.

In conclusion, business analytics represents a transformative approach to managing the complexities of modern business. It enables organizations to exploit the vast potential of data, transforming it into meaningful insights that drive organizational growth and success. Embracing data-driven decision-making allows businesses to remain agile, responsive, and resilient in the face of evolving challenges and opportunities, thus positioning them for sustainable growth and a competitive edge in an ever-changing global market.

1.2 Business Analytics vs. Decision Science

Decision science (in business), or management science, is the application of scientific approaches to make better decisions and solve problems in business and management.

Business analytics and decision science share the common goal of improving decision-making in the business and management context, but they approach this objective from different perspectives and methodologies.

Commonalities:

1. Decision-Making Focus: Both business analytics and decision science center around optimizing decision-making processes within an organization. They aim to enhance the quality of choices made by leveraging data-driven insights and scientific methodologies.
2. Data Utilization: Both disciplines heavily rely on data to support their analyses. Business analytics and decision science draw upon various data sources to inform their decision-making process, ensuring that choices are based on empirical evidence and quantifiable information.
3. Problem-Solving Orientation: Both fields are driven by problem-solving approaches. Whether it's improving operational efficiency, identifying market trends, or enhancing customer experiences, both business analytics and decision science seek to address challenges and capitalize on opportunities through systematic problem-solving.

Differences:

1. Methodological Approach: Business analytics focuses on collecting, analyzing, and interpreting data to generate insights. It relies heavily on statistical techniques, data modeling, and machine learning algorithms to draw meaningful conclusions from data. On the other hand, decision science adopts a more scientific and systematic approach, utilizing mathematical models, optimization techniques, and simulation methods to formulate and evaluate potential solutions.
2. Decision Scope: Business analytics often focuses on analyzing historical data and identifying patterns to inform future decision-making. It primarily deals with descriptive and predictive analytics, aiming to understand what happened and what might happen. In contrast, decision science is more future-oriented, dealing with prescriptive analytics, where the emphasis is on recommending the best course of action to achieve specific objectives.
3. Domain Emphasis: While both disciplines are applicable across various industries and sectors, business analytics is commonly associated with data-intensive fields like marketing, finance, and operations. Decision science, on the other hand, is often associated with areas that require complex optimization and modeling, such as supply chain management, resource allocation, and production planning.

In summary, business analytics and decision science share the common objective of enhancing decision-making in business and management contexts through data-driven insights. While business analytics focuses on analyzing data to draw meaningful insights, decision science adopts a more mathematical and scientific approach to optimize decision-making and solve complex business problems. Both disciplines complement each other, allowing organizations to make well-informed decisions that lead to improved performance and success.

1.3 Why Is Business Analytics Important?

Our definition of business analytics emphasizes facilitation of better and faster decision making in business. More specifically, business analytics enables organizations to make informed, data-driven

decisions that can drive success and competitive advantage. Here are a few reasons why business analytics is important:

1. Insights and Decision Making: Business analytics helps uncover valuable insights from data, allowing organizations to make informed decisions. By analyzing past and current data, organizations can understand customer behavior, market trends, and operational performance, leading to better strategic choices and resource allocation.
2. Performance Optimization: Through business analytics, organizations can identify areas for improvement, optimize processes, and enhance operational efficiency. By analyzing data on key performance indicators (KPIs), organizations can identify bottlenecks, inefficiencies, and opportunities for cost reduction or revenue growth.
3. Predictive Capabilities: Business analytics leverages statistical modeling and predictive analytics to forecast future outcomes. This enables organizations to anticipate market trends, customer preferences, demand fluctuations, and potential risks. By having a forward-looking perspective, organizations can proactively plan and adapt their strategies.
4. Enhanced Customer Understanding: Business analytics helps organizations gain a deeper understanding of their customers. By analyzing customer data, preferences, and behavior, organizations can personalize their offerings, improve customer satisfaction, and develop targeted marketing campaigns. This leads to better customer retention, loyalty, and ultimately, increased profitability.
5. Competitive Advantage: Business analytics provides organizations with a competitive edge. By leveraging data and analytics, organizations can uncover unique insights, identify emerging trends, and capitalize on market opportunities faster than their competitors. It enables organizations to stay agile, adapt to changing market dynamics, and make data-driven decisions that give them a strategic advantage.

1.4 Where Was Business Analytics From?

Business analytics draws upon a wide range of disciplines and sources to derive meaningful insights and drive informed decision-making. These sources include:

1. Statistics: Statistics is fundamental to business analytics as it provides the tools and techniques to analyze and interpret data. Descriptive and inferential statistics help identify patterns, trends, and relationships within the data, enabling data-driven decision-making.
2. Data Mining and Machine Learning: Data mining and machine learning play a crucial role in business analytics by utilizing algorithms and models to uncover hidden patterns and make predictions. They help identify valuable insights from vast and complex datasets, enabling organizations to optimize processes and anticipate future trends.
3. Decision Science/Management Science/Operations Research: Decision science, also known as management science or operations research, uses mathematical and analytical approaches to solve complex business problems. It helps in making optimal decisions related to resource allocation, production planning, and supply chain management, among others.
4. Information Systems and Technology: Business analytics relies on information systems and technology to collect, store, and manage large volumes of data. Advanced tools and technologies facilitate data integration, data cleaning, and data visualization, making it easier to extract insights and communicate findings.
5. Computer Science and Programming: Proficiency in computer science and programming languages is essential for implementing data analytics models and algorithms. Programmers

develop custom solutions and applications to handle specific business challenges and automate analytical processes.

6. Business and Domain Expertise: Business analytics professionals must have a deep understanding of the specific industry or domain they are working in. This domain expertise is crucial for formulating relevant business questions, defining key performance indicators, and interpreting analytical results in the context of the business's goals and objectives.

By drawing upon these diverse sources, business analytics professionals can integrate various disciplines' insights to create a comprehensive and actionable view of the business landscape. This multidisciplinary approach enables organizations to make data-driven decisions that drive innovation, enhance operational efficiency, and achieve sustainable growth.

1.5 Why Business Analytics Now?

Business analytics has gained immense prominence in recent times due to several key factors:

1. More Data: The availability of vast amounts of data, including structured (tabular) and unstructured data, has revolutionized the way businesses operate. Advancements in technology have enabled organizations to capture and store data from various sources, such as social media, online transactions, and sensor networks. This abundance of data provides valuable insights into customer behavior, market trends, and operational processes.
2. Advanced Technologies: The rapid advancements in hardware and software have empowered businesses to handle and process massive datasets efficiently. High-performance computing, cloud computing, and big data infrastructure have made it feasible for organizations to analyze and interpret data on a scale never seen before. These technologies have democratized access to analytics tools, allowing businesses of all sizes to leverage data-driven insights.
3. Sophisticated Models and Algorithms: The development of sophisticated data mining, machine learning, and optimization algorithms has enhanced the accuracy and predictive power of business analytics. These models can identify complex patterns, make accurate predictions, and optimize business processes. Machine learning algorithms, for example, can detect anomalies, classify customer segments, and recommend personalized products or services, driving better decision-making.

In combination, these factors, which were absent a decade or two ago, have created a perfect storm for the adoption of business analytics. Organizations recognize the tremendous potential of data-driven decision-making in gaining a competitive edge. Business analytics empowers businesses to understand customer needs, optimize operations, and seize emerging opportunities promptly. Leveraging these technologies and models helps businesses adapt to dynamic market conditions, identify potential risks, and innovate in an ever-changing landscape. As the volume and variety of data continue to grow, and technologies and models continue to evolve, business analytics will play an increasingly critical role in shaping the future of successful business management.

1.6 Business Analytics vs. Data Analytics vs. Data Science

Business Analytics, Data Analytics, and Data Science are related disciplines with large overlapping areas, and different organizations may use these terms interchangeably or differently based on their context and focus. Here are the commonalities and differences among these fields:

Commonalities:

1. Data Utilization: All three fields involve the collection, processing, and analysis of data to gain insights and make informed decisions.

2. Data-Driven Decision Making: Each field aims to support decision-making processes by providing valuable insights extracted from data analysis.
3. Statistical Techniques: All three disciplines employ statistical methods and techniques to interpret data and draw meaningful conclusions.
4. Focus on Business Impact: Business Analytics, Data Analytics, and Data Science focus on using data to improve business performance, enhance operational efficiency, and drive innovation.

Differences:

1. Scope and Focus: Business Analytics primarily concentrates on using data to address business challenges and improve organizational performance. Data Analytics has a broader focus, encompassing applications in various fields beyond business, such as healthcare, finance, and social sciences. Data Science is even more encompassing, emphasizing the study of data across multiple domains and applications.
2. Data Depth: Data Analytics often deals with structured data in databases and spreadsheets, performing descriptive analysis and visualization. Business Analytics extends this to draw insights specifically relevant to business domains. Data Science involves working with unstructured and diverse data types, utilizing advanced algorithms for predictive modeling and machine learning.
3. Skill Sets: Business Analytics typically requires a strong understanding of business processes and domain expertise. Data Analytics involves proficiency in data manipulation, analysis, and visualization. Data Science demands expertise in programming, machine learning, statistical modeling, and data engineering.
4. Methodology: Business Analytics often centers on generating actionable insights for specific business problems and decision-making scenarios. Data Analytics involves applying statistical and data analysis methods to identify patterns and trends in data. Data Science encompasses a broader approach, including experimental design, data exploration, and hypothesis testing.

Likewise, the job titles "business analyst," "data analyst," and "data scientist" are often utilized with considerable overlap. Job seekers often gravitate towards more contemporary titles, like "data scientist," while hiring organizations may leverage the allure of such titles to attract potential applicants, even if the actual roles and responsibilities align more closely with traditional data analyst or business analyst positions.

In conclusion, while Business Analytics, Data Analytics, and Data Science share common objectives and methods, their scope and application vary. These fields are part of a continuum, with Business Analytics focusing on specific business contexts, Data Analytics embracing broader applications, and Data Science spanning diverse domains with a focus on advanced techniques and research. The varying terminologies used by different organizations can contribute to the fluidity of how these fields are defined and labeled.

1.7 Components of Business Analytics

Business analytics encompasses three key components: descriptive analytics, predictive analytics, and prescriptive analytics. Let's take a quick look at each of them:

1. Descriptive Analytics: Descriptive analytics focuses on understanding historical data and gaining insights into what has happened in the past. It involves summarizing and visualizing data to identify patterns, trends, and key performance indicators (KPIs). Descriptive analytics helps answer questions like "What happened?" and provides a foundation for further analysis.

2. Predictive Analytics: Predictive analytics takes the analysis a step further by using historical data and statistical models to make predictions about future outcomes. By uncovering hidden patterns and relationships within the data, predictive analytics enables organizations to anticipate future trends, behavior, and events. It helps answer questions like "What is likely to happen?" and aids in proactive decision-making.

3. Prescriptive Analytics: Prescriptive analytics goes beyond predicting future outcomes by suggesting the best course of action to achieve desired outcomes. It leverages mathematical optimization, simulation techniques, and decision analysis to provide recommendations and insights on how to optimize decision-making. Prescriptive analytics helps answer questions like "What should we do?" and assists in making informed decisions based on data-driven insights. The overlap between business analytics and decision science predominantly lies in this area.

Together, these three components of business analytics provide a comprehensive framework for extracting valuable insights from data, understanding the past, predicting the future, and guiding decision-making to drive positive business outcomes.

You might come across the term "diagnostic analytics," but in our course, we consider it to be a part of descriptive, predictive, and prescriptive analytics. Therefore, we don't treat it as a separate key component of business analytics.

1.8 What's Covered and Not Covered in the Course?

In this course, we will cover a wide range of topics in Business Analytics, with a focus on applied statistics, data analysis, and data mining techniques. The course aims to equip students with the necessary knowledge and skills to analyze and interpret data effectively for decision-making in business contexts.

1. Applied Statistics to Business Data Analysis:
    - Descriptive Analytics and Data Visualization: We will learn how to explore and visualize data to gain insights into patterns and trends.
    - Basic Probability and Probability Distributions: Students will understand the concepts of probability, conditional probability, statistical independence, and Bayes Rule, which are essential in data analysis.
    - Statistical Inference: Students will learn how to make inferences about population parameters based on sample data, including confidence intervals and hypothesis tests.
2. Basic Data Wrangling:
    - Data Preprocessing and Data Cleaning: We will explore techniques to prepare and clean datasets for analysis, ensuring data quality and consistency.
3. Applied Data Mining and Machine Learning:
    - Supervised Learning: Students will delve into supervised learning techniques, which involve using labeled data to make predictions. Specific methods covered include:
        - Multi-linear Regression: Understanding and applying regression models for predicting numerical outcomes.
        - Logistic Regression and Maximum Likelihood Estimation: Analyzing and predicting binary outcomes.
        - K-Nearest Neighbors (KNN): Employing the KNN algorithm for classification tasks.
        - Naïve Bayes: An introduction to the Naïve Bayes algorithm, a probabilistic classifier.

    - Decision tree: a simple, interpretable, and powerful machine learning model that uses a tree-like structure to make decisions based on input features.
    - Random forest: an ensemble learning method that combines multiple decision trees to improve predictive accuracy and reduce overfitting by averaging their outputs and introducing randomness in the tree-building process.
    - Neural Networks: Exploring basic concepts of artificial neural networks, focusing on multi-layer perceptron model, a powerful machine learning method.
- Unsupervised Learning:
    - Clustering: Students will learn about clustering algorithms to group similar data points based on patterns and characteristics.
    - Association: Understanding association rules and how they identify interesting relationships in large datasets.
- Model Evaluation and Selection:
    - Evaluating the performance of different models and selecting the most appropriate model for specific tasks will be covered.

Throughout the course, students will engage in practical exercises and projects to apply the concepts learned and gain hands-on experience. The course content is designed to provide a comprehensive understanding of essential tools and techniques used in Business Analytics, preparing students to leverage data-driven insights for better decision-making in real-world business scenarios.

While this course primarily focuses on essential topics such as applied statistics, data analysis, and basic data mining and machine learning techniques, it is important to recognize that there are several other critical aspects of the broader field of business analytics not covered in this particular curriculum. Nevertheless, this course provides a solid foundation for understanding the fundamental concepts and methods in business analytics, setting you on the path to explore these advanced topics further in the future. The main topics not covered in this course include the following:

1. Database and SQL: Involves designing and managing databases and using SQL (Structured Query Language) to retrieve and manipulate data efficiently.
2. Data engineering: Focuses on designing and constructing data pipelines, ETL (Extract, Transform, Load) processes, and data infrastructure to handle large-scale data efficiently.
3. Advanced data preprocessing: Explores more sophisticated techniques for data cleaning, imputation, outlier detection, and data transformation.
4. Advanced feature engineering: Involves creating new, meaningful features from existing data to improve model performance and capture important patterns, such as kernel methods.
5. Time series: Focuses on analyzing data collected over time, understanding temporal patterns, and building models suitable for time-dependent data.
6. (Advanced) data mining and machine learning: Covers more complex algorithms and techniques beyond the basic supervised and unsupervised learning methods, such as ensemble methods, deep learning, and reinforcement learning.
7. Optimization: Concentrates on finding the best solution from a set of possible solutions, often used in combination with machine learning for model optimization and parameter tuning.

2. Data Mining vs. Machine Learning

In our course, we interchangeably use the terms 'data mining' and 'machine learning'. Their distinctions are subtle. Some may argue that data mining focuses on discovering patterns, relationships, and insights

from large datasets. It involves extracting valuable information from data to identify hidden trends or knowledge that can be used for decision-making; whereas machine learning focuses on building training algorithms and predictive models and making data-driven predictions or decisions. However, both possesses a defining trait of being driven by data and the techniques employed in both data mining and machine learning substantially overlap, encompassing a spectrum from regression and classification to clustering and association analysis.

In my perspective, machine learning can be considered a modern evolution of data mining, incorporating more recent advancements like reinforcement learning and deep learning. Individuals working with complex neural networks on extensive sets of unstructured data, like language, voice, images, and videos, may find the term "machine learning engineer" more appealing.

2.1 Supervised vs. Unsupervised Learning

Supervised Learning and Unsupervised Learning are two fundamental paradigms within the field of machine learning, each with distinct characteristics and applications.

In supervised learning, the algorithm is trained on a labeled dataset, which means the input data points are paired with corresponding target or output values. The algorithm learns to map inputs to desired outputs based on these labeled examples. The main objective of supervised learning is to learn a mapping function that can accurately predict the output labels for new, unseen data points. Common use cases include classification (where the goal is to assign data points to predefined classes or categories) and regression (where the goal is to predict continuous numerical values). The performance of a supervised learning model is typically evaluated using metrics like accuracy, precision, recall, F1-score (for classification), and Mean Squared Error (for regression).

Unsupervised learning deals with unlabeled data, meaning there are no predefined output labels. The algorithm identifies patterns, structures, or relationships within the data without guidance from labeled examples. The primary goal of unsupervised learning is to discover inherent patterns or groupings in the data without any specific target to predict. Common use cases include clustering (grouping similar data points together) and dimensionality reduction (reducing the number of variables while preserving the underlying structure of the data). Evaluation of unsupervised learning outcomes can be more challenging since there are no predefined labels to compare against. Metrics like silhouette score or visual inspections may be used to assess the quality of clusters or reduced dimensions.

While supervised and unsupervised learning constitute the fundamental paradigms in machine learning, several other paradigms exist. However, this course primarily emphasizes the domain of supervised learning.

2.2 Other Paradigms of Machine Learning

In addition to supervised and unsupervised learning, there are several other paradigms or categories of machine learning that address specific types of tasks and challenges. Some of these paradigms include:

1. **Semi-Supervised Learning:** This paradigm combines elements of both supervised and unsupervised learning. In semi-supervised learning, the training dataset includes a mix of labeled and unlabeled data. The model uses the labeled data for supervised learning tasks and leverages the unlabeled data to capture additional patterns or structures. Semi-supervised learning is particularly useful when obtaining a large amount of labeled data is expensive or time-consuming.

2. **Reinforcement Learning:** Reinforcement learning focuses on training agents to make sequences of decisions in an environment to maximize a reward signal. The agent learns through trial and error, receiving feedback from the environment based on the actions it takes. This paradigm is often used in scenarios where an agent interacts with a dynamic environment, such as game playing, robotics, and autonomous systems.
3. **Transfer Learning:** Transfer learning involves training a model on one task and then applying that knowledge to a different but related task. This approach leverages the learned features from one domain to improve performance in another domain with limited data. Transfer learning is particularly useful when data in the target domain is scarce.
4. **Multi-Task Learning:** In multi-task learning, a single model is trained to perform multiple related tasks simultaneously. The idea is that learning tasks jointly can help improve the performance of each individual task, as the model can leverage shared information and patterns across tasks.
5. **Online Learning:** Online learning, also known as incremental learning or streaming learning, involves updating a model continuously as new data arrives. This is particularly useful in scenarios where data is generated in a stream or when the model needs to adapt to changing patterns over time.
6. **Anomaly Detection:** Anomaly detection focuses on identifying rare or unusual data points that deviate significantly from the norm. This paradigm is widely used in fraud detection, network security, and quality control. Many considers it part of unsupervised learning.

These additional paradigms expand the capabilities of machine learning and address a wide range of real-world problems across different domains and industries. Each paradigm has its own set of techniques, algorithms, and challenges that make it suitable for specific types of tasks.

2.3 Regression vs. Classification

Supervised learning is the primary focus of our course. It entails two fundamental tasks in machine learning: regression and classification, both involving the prediction of outcomes based on input data. However, they serve distinct purposes and are applied in different scenarios.

Regression is a predictive modeling technique used to predict continuous numeric values. In regression, the goal is to establish a relationship between the input variables (also known as features or independent variables) and a target variable (also known as the dependent variable). The objective is to create a function that can accurately estimate the value of the target variable based on the given inputs.

For example, predicting house prices, stock prices, or temperature are common regression tasks. In these cases, the output is a continuous value, and the model aims to find the best-fit line or curve that minimizes the difference between predicted and actual values. Common regression algorithms include Linear Regression, Polynomial Regression, Support Vector Regression, Decision Tree Regression, and Deep Neural Networks for regression.

Classification involves categorizing input data into predefined classes or categories. In this task, the output is a discrete label that represents the class to which the input belongs. The goal of classification is to learn patterns from the input data that can be used to assign the correct class to new, unseen data.

Examples of classification tasks include email spam detection (categorizing emails as spam or not), image recognition (assigning an image to a specific object category), and medical diagnosis (identifying whether a patient has a certain disease or not). Classification algorithms include Logistic Regression, Decision Trees, Random Forest, Support Vector Machines, and Neural Networks.

Throughout our course, we will extensively focus on supervised learning tasks. In the context of the housing price example, the output variable we're interested in is "price." Given that "price" is a continuous variable, this scenario falls under the category of a regression task. On the other hand, in the credit card customer churn example, the target variable (sometimes called 'label') is binary, representing whether a customer churns or not. This classifies the example as a classification task.

In summary, **Regression** predicts continuous numerical values and aims to establish a relationship between input and output. **Classification** assigns data points to discrete categories or classes based on learned patterns. Both regression and classification are essential tools in machine learning, each with its own set of algorithms and techniques tailored to the specific nature of the prediction task.

3. Purpose of Business Analytics: Meaningfulness

Our definition of business analytics alluded that the purpose of business analytics is to transform raw data into ***meaningful*** insights that drive informed decision-making and deliver tangible business value. But what exactly is a meaningful insight? We believe a meaningful insight must have at least the following properties.

1. Relevance: An insight is meaningful if it directly addresses a specific business problem or objective. It should be applicable and provide relevant information that aligns with the goals or challenges of the organization.

2. Actionability: A useful insight should be actionable, meaning it provides practical guidance or recommendations for decision-making.

3. Impact: An insight is meaningful if it has the potential to make a significant impact on business outcomes. It should have the capacity to drive positive changes, improve efficiency, enhance performance, increase revenue, reduce costs, or mitigate risks.

4. Credibility: A meaningful insight should be based on reliable data and robust analysis. It should be supported by evidence, statistical significance, or a sound logical foundation to ensure its credibility.

5. Measurability or quantifiability: It refers to the ability to quantify the results or outcomes derived from data analysis. In other words, the insights obtained should be expressed in measurable terms, allowing for precise and quantitative evaluation of their impact or significance. Measurability ensures that the insights can be translated into actionable metrics or key performance indicators (KPIs), enabling businesses to monitor progress and make data-driven decisions. A meaningful business analytics insight should provide clear and specific numerical values, percentages, or trends, enabling stakeholders to understand the magnitude of the findings. Without measurability, insights become ambiguous or subjective, making it challenging to assess their actual impact on business outcomes. By focusing on measurable insights, organizations can set specific targets, benchmark their performance, and evaluate the effectiveness of strategies or interventions.

6. The generalizability of insight refers to the extent to which the knowledge or findings derived from analyzing a specific dataset can be applied or generalized to a broader context or

population. In other words, it assesses the validity and applicability of an insight beyond the immediate data used for its discovery.

Data or datasets used for analysis are almost always gathered from historical records and represent a subset of the larger population of interest. While analyzing this data can provide valuable insights, it is important to ensure that these insights are applicable to the broader population and have relevance for future decision-making. Ultimately, when making business decisions, the focus is on the present and future rather than past; on the entire population under consideration such as all customers rather than a subset of it.

3.1 Temporal vs. Cross-Sectional Generalizability

Generalizability is an ultimate measurement of meaningfulness of a business analytics project. The data collected are from the past while the problems we want to solve lie in the present or future. Even with streaming data, there exists a time latency or delay. Business environments are dynamic, and conditions can evolve over time. Therefore, it is crucial to consider the temporal aspect when evaluating the generalizability of insights. The Greek philosopher Heraclitus famously stated, "You cannot step into the same river twice."

Temporal generalizability refers to the extent to which insights or findings derived from analyzing historical data remain valid and applicable over time. It assesses the ability of insights to withstand changes and trends that occur in the future. Naturally, data collected over different time periods are called time-series data.

It is important to note that temporal generalizability does not imply that insights will remain static or unchanging. Rather, it emphasizes the ability of insights to capture enduring trends and relationships that are likely to persist over time. However, businesses should continuously monitor and validate insights against new data and evolving circumstances to ensure their ongoing relevance and applicability.

Achieving temporal generalizability poses a significant challenge in the field of data analysis. This challenge arises due to the inherent nature of data or datasets, which are often collected from the past and represent a subset of a larger population of interest. Temporal generalizability generally deteriorates as time goes by. Similar to weather forecasting, making accurate predictions becomes increasingly challenging the further into the future we look. The reason for this lies in the dynamic and complex nature of the business landscape.

Over time, various external factors and internal dynamics can significantly impact the conditions under which businesses operate. Market trends shift, consumer preferences evolve, new technologies emerge, and regulatory frameworks change. These ongoing changes introduce uncertainties that can diminish the reliability of insights and predictions.

In the context of data analysis, temporal generalizability relies on the assumption that historical patterns and relationships will persist in the future. However, as time progresses, the accuracy of this assumption weakens. External disruptions, unforeseen events, and unanticipated shifts in the business environment can alter the underlying dynamics, rendering past insights less applicable to future scenarios.

Moreover, the accumulation of new data over time introduces additional complexities. New data points may deviate from previous patterns, making it necessary to reassess and update models regularly. As

more time elapses, the potential for significant deviations and unforeseen outliers increases, further challenging the reliability and generalizability of insights.

To mitigate the diminishing temporal generalizability, analysts employ various techniques. These may include incorporating leading indicators, continuously monitoring and updating models, and adopting more dynamic and adaptable forecasting approaches. By acknowledging the inherent limitations of long-term predictions and actively adapting to changing conditions, analysts can enhance the temporal generalizability of their insights, albeit with a recognition that future outcomes become increasingly uncertain as time progresses.

The counterpart of temporal generalizability is cross-sectional generalizability. While temporal generalizability focuses on the applicability of insights over time, cross-sectional generalizability refers to the extent to which insights derived from analyzing a specific sample or dataset can be generalized to a broader population or different settings at a given point in time.

Cross-sectional generalizability involves assessing the validity and relevance of insights across different groups or contexts that share similar characteristics or features. It ensures that the insights obtained from a particular dataset or sample are representative and applicable to a wider population or similar situations. Data collected at the same point in time are called cross-sectional data.

In the context of cross-sectional generalizability, we encounter two key statistics concepts: population and sample. The population represents the entire group of interest, consisting of all elements or individuals we aim to draw conclusions about. On the other hand, a sample is a subset of the population that we collect data from to make inferences and predictions about the larger group. The challenge lies in ensuring that the findings from our sample are applicable and generalizable to the entire population, allowing us to draw meaningful insights that can be useful in broader contexts.

To evaluate cross-sectional generalizability, considerations include the representativeness and diversity of the sample, the relevance of the analyzed variables or factors, and the consistency of the relationships or patterns observed across different subsets or populations. Statistical techniques, such as random sampling or stratified sampling, can help enhance cross-sectional generalizability by ensuring a more representative selection of data.

By considering both temporal generalizability and cross-sectional generalizability, businesses can obtain a comprehensive understanding of the validity and applicability of insights. This allows for informed decision-making that takes into account both the dynamics of time and the broader population or contexts in which the insights are intended to be applied.

3.2 Measuring Generalizability

Measuring generalizability in the context of business analytics involves assessing how well a model performs on unseen data to ensure that it can be applied effectively to new, real-world situations. One common approach to measuring generalizability is through dataset splitting and evaluating model performance on a separate test set.

The process starts by dividing the available dataset into three distinct subsets: the training set, the validation set, and the test set. The training set is used to build the model, the validation set is employed to fine-tune hyperparameters and prevent overfitting, and the test set remains untouched until the final evaluation.

Once the model is trained and validated using the training and validation sets, it is evaluated on the test set. The performance metrics obtained on the test set serve as an estimate of how well the model will perform on new, unseen data. It helps gauge the generalizability of the model beyond the data it was trained on.

If the model performs well on the test set, it indicates that it has generalizability and can make reliable predictions on new data. On the other hand, if there is a significant drop in performance between the training and test sets, it suggests that the model is overfitting and may not perform well on new data. By following this process of dataset splitting and evaluating model performance on the test set, we can ensure that the business analytics model is robust and reliable for future decision-making tasks.

3.3 Overfitting vs. underfitting

Underfitting and overfitting are two common issues that can occur when building predictive models in data analytics and machine learning:

1. Underfitting: Underfitting occurs when a model is too simple to capture the underlying patterns in the data. As a result, it performs poorly on both the training data and new, unseen data. Essentially, the model fails to learn from the data and lacks the ability to generalize to new situations. Underfitting can be recognized when the model's performance is consistently low and doesn't improve with more training data or complex models.

2. Overfitting: Overfitting happens when a model becomes overly complex and starts memorizing noise or random fluctuations in the training data rather than learning meaningful patterns. As a consequence, the model performs exceptionally well on the training data but fails to generalize to new data. One can observe overfitting when the model's performance on the training data is high, but it drops significantly when evaluated on new, unseen data.

To address these issues, finding the right balance between simplicity and complexity is crucial. Underfitting can be mitigated by using more complex models, increasing model complexity, or introducing more relevant features. On the other hand, overfitting can be reduced by using simpler models, reducing the number of features, and employing techniques like cross-validation and regularization. The goal is to find a model that fits the data well while maintaining the ability to generalize to new, unseen data.

4. Process of Business Analytics Project

The process of a business analytics project typically involves several key steps. While the specific details may vary depending on the project and organization, here is a general outline of the process:

1. Define the Problem: Clearly understand the business problem or objective that the analytics project aims to address. Identify the key questions to be answered or decisions to be supported through analytics.
2. Data Collection: Gather relevant data from various sources, ensuring its quality, completeness, and accuracy. This may involve accessing databases, acquiring external data, or conducting surveys.

3. Data Cleaning and Preparation: Clean and preprocess the data to remove errors, inconsistencies, and missing values. Transform and structure the data in a way that is suitable for analysis.
4. Exploratory Data Analysis: Explore and visualize the data to gain insights, identify patterns, detect outliers, and understand the relationships between variables. This step helps uncover initial trends and potential areas of focus.
5. Data Modeling and Analysis: Apply appropriate statistical, data mining, or machine learning techniques to build models and analyze the data. This involves selecting the right algorithms, training models, and evaluating their performance. Data preprocessing and feature engineering are included in the data modeling step to enhance the quality of the data and create relevant features that can improve the performance of machine learning models, leading to more accurate and meaningful insights.
6. Interpretation and Insights: Interpret the results obtained from the analysis and translate them into meaningful insights and actionable recommendations. Connect the findings to the initial business problem and address the questions or decisions identified in the beginning.
7. Reporting and Visualization: Present the findings in a clear, concise, and visually appealing manner. Use charts, graphs, and other visualizations to effectively communicate the insights and recommendations to stakeholders.
8. Implementation and Monitoring: Collaborate with relevant stakeholders to implement the recommended actions or strategies based on the insights gained. Establish monitoring mechanisms to track the impact and effectiveness of the implemented solutions.
9. Continuous Improvement: Analyze the outcomes and learn from the project. Evaluate the success of the analytics project, identify areas for improvement, and incorporate feedback to refine future projects or iterations.

Throughout the process, effective communication, collaboration, and engagement with stakeholders are crucial. It is important to involve domain experts, business leaders, and data analysts or scientists to ensure that the analytics project aligns with the organization's goals and generates valuable outcomes.

Keep in mind that this is a general framework, and the complexity and duration of each step may vary depending on the project's scope, data availability, and specific requirements.

In our course, we put a strong emphasis on steps 4, 5, and 6 of the analytics process. We also cover steps 3 and 7 to some extent.

5. Define a Business Analytics Problem

In practice, defining a business analytics problem or project involves several key steps to ensure clarity, relevance, and actionable outcomes:

- Identify the Business Objective: Clearly articulate the specific business objective or question that the analytics project aims to address. The objective should align with the organization's overall goals and priorities.
- Understand Stakeholder Needs: Engage with stakeholders, including management, clients, or end-users, to gather their requirements, expectations, and pain points. Understanding stakeholder needs ensures that the project addresses the most critical business challenges.

- Formulate the Research Question: Translate the business objective into a well-defined research question or hypothesis that can be tested using data analysis. The question should be specific, measurable, and relevant to the business context.
- Define Key Performance Indicators (KPIs): Establish clear and quantifiable metrics to measure the success of the analytics project. KPIs should align with the business objective and provide tangible indicators of project outcomes.

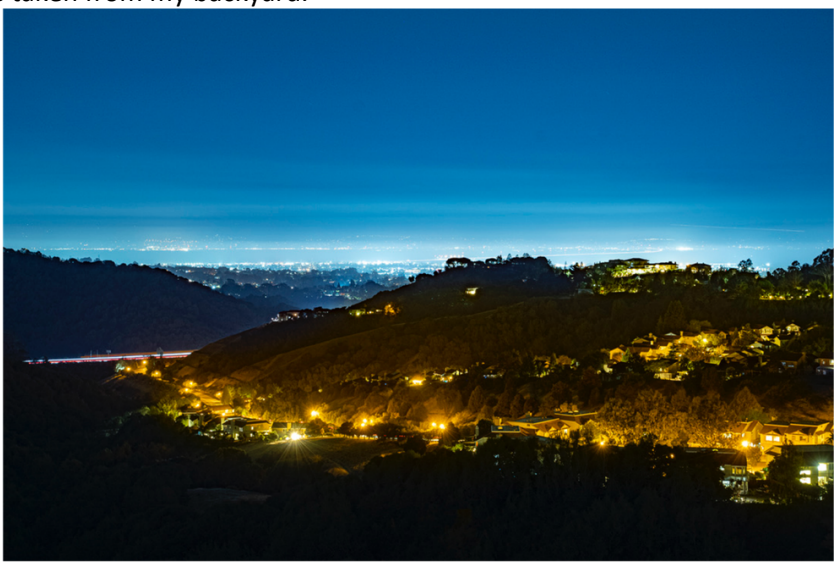
The definition of a business analytics problem is a crucial and foundational step that sets the trajectory for the entire analytics project. By clearly articulating the specific business objective or question to be addressed, organizations can effectively determine the subsequent key elements of the project:

- Data Requirements: Defining the business analytics problem helps identify the necessary data sources and variables needed to address the research question. Understanding the type, volume, and quality of data required ensures that relevant and reliable information is available for analysis.
- Analytical Methods: The nature of the business analytics problem guides the selection of appropriate analytical methods and techniques. Whether it involves statistical analysis, data mining, machine learning, or optimization algorithms, choosing the right methods ensures that the insights generated are aligned with the research question and data characteristics.
- Scope and Feasibility: The definition of the problem establishes the scope of the analytics project, considering the time, resources, and data availability. This step helps organizations assess the feasibility of achieving the project objectives within the given constraints.
- Potential Contributions: Clearly defining the business analytics problem also allows stakeholders to understand the potential contributions of the project. It highlights the value that insights and actionable recommendations derived from data analysis can bring to the organization in terms of improved decision-making, enhanced performance, and strategic advantages.

Overall, a well-defined business analytics problem serves as the compass that guides the project, ensuring that efforts are focused on addressing the most relevant business challenges and driving positive outcomes. It lays the groundwork for a successful analytics journey, leading to valuable insights, informed decision-making, and tangible benefits for the organization.

5.1 My Motivation and Project Definition

Below is a picture taken from my backyard.



During a house hunting trip years ago, we had the opportunity to explore several houses along a street. Notably, one house on the street boasted a magnificent view, as shown in the provided picture. On the other side of the street, two houses were remarkably similar, except that they lacked any view. The house with the splendid view was listed at a significantly higher price, and despite uncertainties, we eventually decided to purchase this house. However, the question of whether we potentially overpaid for the view has lingered, given the absence of concrete data at the time.

Years later, I stumbled upon a dataset that offers an opportunity to quantitatively assess the value of a view in relation to house prices. Consequently, my project aims to comprehensively understand the impact of a view on house prices, both in terms of dollar amounts and as a percentage of the overall house value. Through detailed data analysis and modeling, I intend to gain valuable insights into the extent to which a view influences housing price.
Studying the effect of view on housing prices is of significant importance as it can provide valuable insights for both homeowners and real estate investors. Understanding how the presence of a view impacts housing prices can inform decisions related to property valuation, investment opportunities, and overall property desirability.

The contribution of this study lies in its ability to quantitatively evaluate the influence of a view on housing prices, providing concrete data-driven evidence to support decision-making. By analyzing the relationship between view and housing prices, we can gain a deeper understanding of the specific value a view adds to a property, both in terms of dollar amounts and as a percentage of the overall house value.

Moreover, the methodology used in this study can be widely applied to investigate similar problems in the real estate domain and beyond. By employing statistical analysis, regression modeling, and data-driven approaches, researchers can explore various factors that influence property prices, such as location, amenities, or other external factors.

The insights gained from this study can be invaluable for homeowners, buyers, and investors, helping them make more informed decisions when evaluating properties and negotiating prices. Furthermore, the generalizable nature of the methodology enables its application to study various aspects of the real estate market and other domains where quantitative assessment of influential factors is crucial.