

Prepared by Dr. Wesam Ahmed

IT393
Business Intelligence

Lecture 10



Prescriptive Analytics and Big Data Concepts and Tools

Lecture 10

IT393

Business Intelligence

Prepared by
Dr. Wesam Ahmed

01

Learning Objectives:

- 1.1 Understand the applications of prescriptive analytics techniques
- 1.2 Understand the basic concepts of analytical decision modeling
- 1.3 Learn what Big Data is and how it is changing the world of analytics
- 1.4 Understand the motivation for Big Data analytics
- 1.5 The wide range of enabling technologies for Big Data analytics
- 1.6 Case study about data mining

02

Prescriptive analytics

Prescriptive analytics is the process of using data to determine an optimal course of action. By considering all relevant factors, this type of analysis yields recommendations for next steps. Because of this, prescriptive analytics is a valuable tool for data-driven decision-making.

Most organizations face the problem of **making decisions where one has to select from multiple options**. Each option has a cost and capability associated with it. **The goal of such models** is to select **the combination of options** that meet all the requirements and yet optimizes the costs. **Prescriptive analytics** particularly **apply to the problem of such decisions**. And tools such as **built-in or Premium Solver for Excel** make it easy to apply such techniques.

03

Prescriptive analytics

Prescriptive Analytics Model Examples: Modeling is a key element for prescriptive analytics. In the examples, one has to employ a mathematical model to be able to recommend a decision for any realistic problem. For example, deciding which customers (among potentially millions) will receive what offer so as to maximize the overall response value but staying within a budget is not something you can do manually. Building a probability-based response maximization model with the budget as a constraint would give us the information we are seeking. Depending on the problem we are addressing, there are many classes of models, and there are often many specialized techniques for solving each one.

04

Prescriptive analytics

Category	Process and Objective	Representative Techniques
Optimization of problems with few alternatives	Find the best solution from a small number of alternatives	Decision tables, decision trees, analytic hierarchy process
Optimization via algorithm	Find the best solution from a large number of alternatives, using a step-by-step improvement process	Linear and other mathematical programming models, network models
Optimization via an analytic formula	Find the best solution in one step, using a formula	Some inventory models
Simulation	Find a good enough solution or the best among the alternatives checked, using experimentation	Several types of simulation
Heuristics	Find a good enough solution, using rules	Heuristic programming, expert systems
Predictive models	Predict the future for a given scenario	Forecasting models, Markov analysis
Other models	Solve a what-if case, using a formula	Financial modeling, waiting lines

TABLE 10.1:
Categories of
Models

05

Prescriptive analytics

How Prescriptive Analytics Works :

Prescriptive analytics tries to answer the question “How do we get to this point?” It relies on artificial intelligence (AI) techniques, such as machine learning (the ability of a computer program without additional human input), to understand and advance from the data it acquires, adapting all the while.

Machine learning makes it possible to process a tremendous amount of data available today. As new or additional data becomes available, computer programs adjust automatically to make use of it, in a process that is much faster and more comprehensive than human capabilities could manage.

Prescriptive analytics works with another type of data analytics: predictive analytics, which involves the use of statistics and modeling to determine future performance, based on current and historical data.

06

Prescriptive analytics

Structure of Mathematical Models for Decision Support:

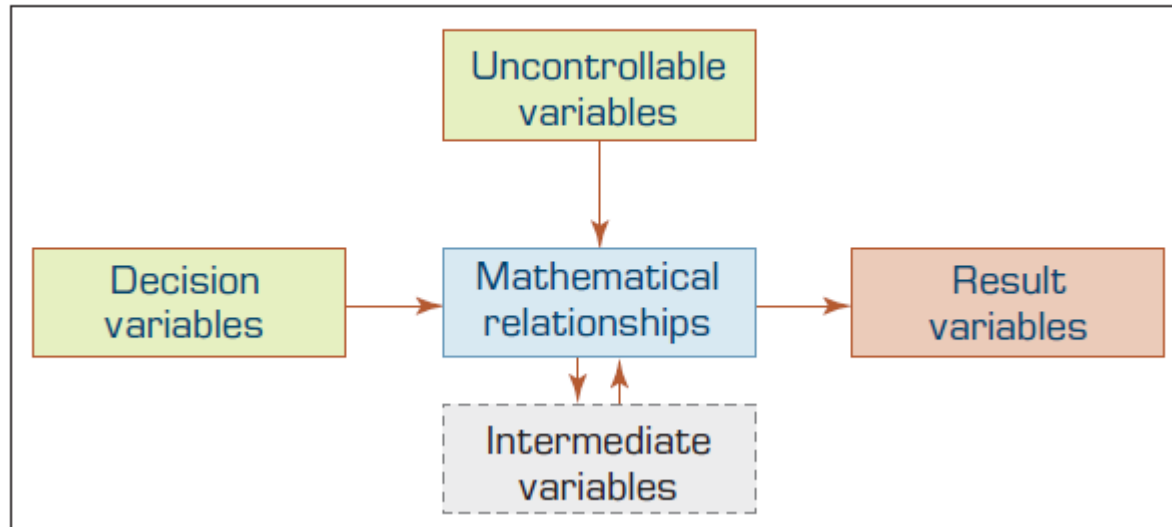


Figure 10.1 The General Structure of a Quantitative Model.

07

Prescriptive analytics

A quantitative decision support model typically has **four basic components** that are connected through mathematical relationships:

- **Result (Outcome) Variables**
- **Decision Variables**
- **Uncontrollable Variables (Parameters)**
- **Intermediate Result Variables**

The model's purpose is to determine how decisions and external factors influence outcomes.

08

Prescriptive analytics**Result (Outcome) Variables** (Outputs / Dependent Variables)

- Show how well the system meets its goals (effectiveness).
- Dependent on decision variables and uncontrollable variables.
- Example: Total profit, customer satisfaction, production output.

Decision Variables (Controllable Inputs)

- Chosen by the decision maker; represent possible courses of action.
- Example: Investment amount, scheduling choices, pricing.

09

Prescriptive analytics**Uncontrollable Variables (Parameters) (External Constraints)**

Factors outside the decision maker's control that affect results.

May be fixed (parameters) or variable.

Example: Interest rates, tax regulations, building codes.

Intermediate Result Variables (Linking Variables)

Represent intermediate outcomes that influence final results.

Example: Employee salaries → affect satisfaction → affect productivity.

10

Prescriptive analytics

The Structure of Mathematical Models : The components of a quantitative model are linked by mathematical (algebraic) expressions—equations or inequalities. A very simple financial model is

$$P = R - C$$

where P = profit, R = revenue, and C = cost. This equation describes the relationship among the variables. Another well-known financial model is the simple present-value cash flow model, where P = present value, F = a future single payment in dollars, i = interest rate (percentage), and n = number of years. With this model, it is possible to determine the present value of a payment of \$100,000 to be made 5 years from today, at a 10% (0.1) interest rate, as follows:

$$P = 100,000 / (1 + 0.1)^5 = 62,092$$

11

Prescriptive analytics

Examples of Prescriptive Analytics

Numerous data-intensive businesses and government agencies can benefit from using prescriptive analytics. This includes companies in the financial services and healthcare [sectors](#), where the cost of human error is high. For instance, **prescriptive analytics could be used to:**

- Predict whether an article on a particular topic will be popular with readers based on data about searches and social shares for related topics
- Adjust a worker training program in real time based on how the worker is responding to each lesson.

12

Prescriptive analytics

Banking: Fraud Detection : The use of prescriptive analytics is the **detection and flagging of bank fraud**. With the sheer volume of data stored in a bank's system, it would be nearly impossible for a person to manually detect any suspicious activity in a single account. An algorithm—trained using customers' historical transaction data—analyzes and scans new transactional data for anomalies. **For instance, perhaps you typically spend \$3,000 per month, but this month, there's a \$30,000 charge on your credit card.**

The algorithm analyzes patterns in your transactional data, alerts the bank, and provides a recommended course of action. In this example, the course of action may be to cancel the credit card, as it could have been stolen.

13

Prescriptive analytics

Product Management: Development and Improvement : Prescriptive analytics can also inform product development and improvements. Product managers can gather user data by surveying customers, running tests with a product's beta versions, conducting market research with people who aren't current product users, and collecting behavioral data as current users interact. All this data can be analyzed—either manually or algorithmically—to identify trends, discover the reasons for those trends, and predict whether the trends are predicted to recur.

Prescriptive analytics can help determine which features to include or leave out of a product and what needs to change to ensure an optimal user experience.

14

Prescriptive analytics

Leveraging Prescriptive Analytics at Your Organization: If your organization is new to prescriptive analytics, there's no better time to see how it impacts your decision-making processes. **Start small with one question you need answered or one process you'd like to optimize.** Gather data surrounding that question or process and move through each type of analytics to paint the full picture. **Descriptive:** What trends does the data show?

- **Diagnostic:** What factors contribute to those trends? Why are those trends occurring?
- **Prescriptive:** If applicable, determine whether a trend is one you can expect to continue or recur.
- **Prescriptive:** Finally, dive into prescriptive analytics. If you have a proprietary algorithm or third-party analytics tool, run it using your company's data. Alternatively, conduct manual analysis of possible next steps based on what you've discovered about your question or process. How will each option impact the situation's outcome and, thus, your goal?

15

Big data

Big Data, which means **many things to many people**, is not a new technological fad. It has become a business priority that has the potential to profoundly change the competitive landscape in today's globally integrated economy. In addition to **providing innovative solutions to enduring business challenges**, **Big Data and analytics instigate new ways to transform processes, organizations, entire industries, and even society altogether.**

Using data to understand customers/clients and business operations to sustain growth and profitability is an increasingly challenging task for today's enterprises. **As more and more data becomes available in various forms and fashions, timely processing of the data with traditional means becomes impractical.** Nowadays, this phenomenon called Big Data, which is receiving substantial press coverage and drawing increasing interest from both business users and IT professionals. **The result is that Big Data is becoming an overhyped and overused marketing buzzword.**

16

Big data

the term **Big Data** has been used to describe the massive volumes of data analyzed by huge organizations like Google or research science projects at NASA. But for most businesses, it's a relative term: "Big" depends on an organization's size. Big Data exceeds the reach of commonly used hardware environments and/or capabilities of software tools to capture, manage, and process it within a tolerable time span for its user population.

Big Data has become a popular term to describe the exponential growth, availability, and use of information, both structured and unstructured. Because of the technology challenges in managing the large volume of data coming from multiple sources, sometimes at a rapid speed, additional new technologies have been developed to overcome the technology challenges. Use of the term **Big Data** is usually associated with such technologies. Because a prime use of storing such data is generating insights through analytics, sometimes the term Big Data is expanded as Big Data analytics.

17

Big data

Big Data may come from Web logs, radio-frequency identification (RFID), global positioning systems (GPS), sensor networks, social networks, Internet-based text documents, Internet search indexes, detail call records, astronomy, atmospheric science, biology, genomics, nuclear physics, biochemical experiments, medical records, scientific research, military surveillance, photography archives, video archives, and large-scale e-commerce practices. **Big Data is typically defined by five “V”s:**

- **Velocity** is the speed at which the data is created and how fast it moves.
- **Volume** is the amount of data qualifying as big data.
- **Value** is the value the data provides.
- **Variety** is the diversity that exists in the types of data.
- **Veracity** is the data's quality and accuracy.

18

Big data

Velocity: refers to how quickly data is generated and how fast it moves. This is an important aspect for organizations that need their data to flow quickly, so it's available at the right times to make the best business decisions possible. An organization that uses big data will have a large and continuous flow of data that's being created and sent to its end destination. Data could flow from sources such as machines, networks, smartphones or social media. Velocity applies to the speed at which this information arrives.

For example, in healthcare, many medical devices today are designed to monitor patients and collect data. From in-hospital medical equipment to wearable devices, collected data needs to be sent to its destination and analyzed quickly.

19

Big data

Volume refers to **the amount of data that exists**. Volume is like the base of big data, as it's the initial size and amount of data that's collected. **If the volume of data is large enough, it can be considered big data**. However, what's considered to be big data is relative and will change depending on the available computing power that's on the market.

For example, a company that operates hundreds of stores across several states **generates millions of transactions per day**. This qualifies as big data, and the average number of total transactions per day across stores represents its volume.

20

Big data

Value refers to the benefits that big data can provide, and it relates directly to what organizations can do with that collected data. Being able to pull value from big data is a requirement, as the value of big data increases significantly depending on the insights that can be gained from it.

Organizations can use big data tools to gather and analyze the data, but how they derive value from that data should be unique to them. Tools like Apache Hadoop can help organizations store, clean and rapidly process this massive amount of data.

A great example of big data value can be found in the gathering of individual customer data. When a company can profile its customers, it can personalize their experience in marketing and sales, improving the efficiency of contacts and garnering greater customer satisfaction.

21

Big data

Variety refers to the diversity of data types. An organization might obtain data from several data sources, which might vary in value. Data can come from sources in and outside an enterprise as well. The challenge in variety concerns the standardization and distribution of all data being collected. Collected data can be unstructured, semi-structured or structured.

Raw data also qualifies as a data type. While raw data can fall into other categories -- structured, semi-structured or unstructured -- it's considered *raw* if it has received no processing at all. Most often, raw applies to data imported from other organizations or submitted or entered by users. Social media data often falls into this category.

A more specific example could be found in a company that gathers a variety of data about its customers. This can include structured data culled from transactions or unstructured social media posts and call center text. Much of this might arrive in the form of raw data, requiring cleaning before processing.

22

Big data

Veracity refers to the quality, accuracy, integrity and credibility of data. Gathered data could have missing pieces, might be inaccurate or might not be able to provide real, valuable insight. Veracity, overall, refers to the level of trust there is in the collected data. Data can sometimes become messy and difficult to use. A large amount of data can cause more confusion than insights if it's incomplete. For example, in the medical field, if data about what drugs a patient is taking is incomplete, the patient's life could be endangered. Both value and veracity help define the quality and insights gathered from data.

Variability refers to inconsistencies in either the usage or the flow of big data. Minimizing variability in big data requires carefully constructing data flows as data moves through the organization's systems, from the transactional to the analytical and everything in between. The biggest benefit is big data veracity, as consistency in data usage yields more stable reporting and analytics and therefore higher confidence.

23

Big data

Business Problems Addressed by Big Data Analytics

The top business problems addressed by Big Data overall are **process efficiency and cost reduction**, as well as **enhancing customer experience**, but different priorities emerge when it is looked at by industry. **Process efficiency and cost reduction** are perhaps among the top-ranked problems that can be addressed with **Big Data analytics** for the manufacturing, government, energy and utilities, communications and media, transport, and healthcare sectors. **Enhanced customer experience** may be at the top of the list of problems addressed by insurance companies and retailers. Here is a partial list of problems that can be addressed using Big Data analytics:

24

Big data

List of Problems Big Data Can Address:

- **Process efficiency and cost reduction**
 - Streamlining operations, reducing waste, cutting costs.
- **Brand management**
 - Monitoring brand reputation through social media and customer sentiment.
- **Revenue maximization, cross-selling, and up-selling**
 - Using data to find ways to increase sales to existing customers.
- **Enhanced customer experience**
 - Personalizing services, improving product recommendations.

25

Big data

- **Identifying new products and market opportunities**
 - Detecting gaps in the market before competitors do.
- **Risk management**
 - Predicting and mitigating risks (fraud detection, investment risks, etc.).
- **Regulatory compliance**
 - Using data to ensure laws and rules are followed.
- **Enhanced security capabilities**
 - Detecting cyberattacks or security breaches in real time.
- **Churn identification & customer recruiting** : Spotting customers likely to leave, and attracting new ones.
- **Improved customer service** : Predicting issues and responding quickly.

26

Big data

Big Data Technologies:

There are a number of technologies for **processing and analyzing Big Data**, but most have some common characteristics . Namely, they take advantage of commodity hardware to enable scale-out and **parallel-processing techniques**; employ nonrelational data storage capabilities to **process unstructured and semistructured data**; and apply **advanced analytics and data visualization technology to Big Data** to convey insights to end users. The three Big Data technologies that stand out that most believe will transform the business analytics and data management markets are **MapReduce, and Hadoop**.

MapReduce is a technique **popularized by Google** that distributes the **processing of very large multistructured data files** across a large cluster of machines. **High performance** is achieved by **breaking the processing into small units of work** that can be **run in parallel** across the hundreds, potentially thousands, of nodes in the cluster.

27

Big data

MapReduce is a **programming model** and an associated implementation for processing and generating large data sets. Programs written in this functional style are automatically **parallelized and executed on a large cluster of commodity machines**. This **allows programmers without any experience** with parallel and distributed systems to easily utilize the resources of a large distributed system.

Hadoop is an **open source framework** for **processing, storing, and analyzing massive amounts** of distributed, unstructured data. Originally created by Doug Cutting at Yahoo!, **Hadoop was inspired by MapReduce**, a user-defined function developed by Google in the early 2000s for indexing the Web. **It was designed to handle petabytes and exabytes of data distributed over multiple nodes in parallel**. Hadoop breaks up Big Data into multiple parts so each part can be processed and analyzed at the same time.

<https://www.kaggle.com/datasets/adilshamim8/exploring-mental-health-data/data>

About Dataset

This dataset presents a synthetic yet highly realistic representation of mental health survey responses, meticulously crafted using a **deep learning generative model** trained on the original *Depression Survey/Dataset for Analysis*. It serves as the foundation for the Playground Series - Season 4, Episode 11 competition.

Designed to simulate complex mental health assessment patterns while preserving participant anonymity, this dataset invites exploration into psychological data modeling, ethical AI applications, and modern classification techniques.

Dataset Files

- `train.csv` : Contains anonymized survey features with a **binary target variable** `class` , labeled as:
 - `e` : One class of mental health profile
 - `p` : Another class (labels intentionally obfuscated to promote unbiased modeling)
- `test.csv` : Contains the same features as the training set, excluding the target. Used for prediction.

Column Descriptions

Columns: 19

- **id** : Unique identifier for each participant record.
- **Name** : Placeholder for the participant's name (anonymized or synthetic).
- **Gender** : Reported gender of the participant (e.g., Male, Female, Other).
- **Age** : Age of the respondent, typically in years.
- **City** : City or region of residence, possibly synthetic.
- **Working Professional or Student** : Indicates whether the respondent is currently a student or a working professional.
- **Profession** : Declared field of profession or study (e.g., Engineering, Business).
- **Academic Pressure** : Self-reported academic stress level (e.g., High, Moderate, Low).
- **Work Pressure** : Self-reported job-related stress level (for working professionals).

30

Case study of data mining

- **CGPA** : Cumulative Grade Point Average—numerical academic performance indicator.
- **Study Satisfaction** : Subjective measure of satisfaction with current study routine.
- **Job Satisfaction** : Subjective measure of satisfaction with current job role (for professionals).
- **Sleep Duration** : Average hours of sleep per day.
- **Dietary Habits** : Reported diet quality or eating patterns (e.g., Balanced, Skipped meals).
- **Degree** : Academic degree currently pursued or achieved (e.g., Diploma, Bachelor's, Master's).
- **Have you ever had suicidal thoughts ?** : Binary or categorical response indicating prior mental health challenges.
- **Work/Study Hours** : Average number of hours spent working or studying per day.
- **Financial Stress** : Level of financial burden or concern (e.g., Severe, Mild, None).
- **Family History of Mental Illness** : Whether the respondent has a known family history of mental health issues.

31

Case study of data mining

```
df=pd.read_csv('/kaggle/input/exploring-mental-health-data/test.csv')
df
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 93800 entries, 0 to 93799
Data columns (total 19 columns):
#   Column                                     Non-Null Count  Dtype
---  -
0    id                                         93800 non-null  int64
1    Name                                       93800 non-null  object
2    Gender                                    93800 non-null  object
3    Age                                        93800 non-null  float64
4    City                                       93800 non-null  object
5    Working Professional or Student          93800 non-null  object
6    Profession                               69168 non-null  object
7    Academic Pressure                        18767 non-null  float64
8    Work Pressure                            75022 non-null  float64
9    CGPA                                      18766 non-null  float64
10   Study Satisfaction                       18767 non-null  float64
11   Job Satisfaction                         75026 non-null  float64
12   Sleep Duration                           93800 non-null  object
13   Dietary Habits                           93795 non-null  object
14   Degree                                   93798 non-null  object
15   Have you ever had suicidal thoughts ?    93800 non-null  object
16   Work/Study Hours                        93800 non-null  float64
17   Financial Stress                        93800 non-null  float64
18   Family History of Mental Illness         93800 non-null  object
dtypes: float64(8), int64(1), object(10)
memory usage: 13.6+ MB
```

32

Case study of data mining

```
df.columns
```

```
Index(['id', 'Name', 'Gender', 'Age', 'City',  
      'Working Professional or Student', 'Profession', 'Academic Pressure',  
      'Work Pressure', 'CGPA', 'Study Satisfaction', 'Job Satisfaction',  
      'Sleep Duration', 'Dietary Habits', 'Degree',  
      'Have you ever had suicidal thoughts ?', 'Work/Study Hours',  
      'Financial Stress', 'Family History of Mental Illness'],  
      dtype='object')
```

32

Case study of data mining

```
df.isnull().sum()
```

```
id          0
Name        0
Gender      0
Age         0
City        0
Working Professional or Student  0
Profession  24632
Academic Pressure  75033
Work Pressure  18778
CGPA        75034
Study Satisfaction  75033
Job Satisfaction  18774
Sleep Duration  0
Dietary Habits  5
Degree      2
Have you ever had suicidal thoughts ?  0
Work/Study Hours  0
Financial Stress  0
Family History of Mental Illness  0
dtype: int64
```

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

# Load dataset
df = pd.read_csv('/kaggle/input/exploring-mental-health-data/test.csv')

# Select numerical columns
numerical_df = df.select_dtypes(include=['int64', 'float64'])

# Compute correlation matrix
corr = numerical_df.corr()

# Plot heatmap
plt.figure(figsize=(10, 8))
sns.heatmap(corr, annot=True, cmap='coolwarm', fmt=".2f", linewidths=0.5, square=True)
plt.title("Correlation Heatmap of Numerical Features", fontsize=16)
plt.tight_layout()
plt.show()
```

34

Case study of data mining

```
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder, StandardScaler
from sklearn.impute import SimpleImputer
from sklearn.metrics import classification_report, confusion_matrix, roc_curve, auc, RocCurveDisplay
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from xgboost import XGBClassifier
from sklearn.svm import SVC
from sklearn.neighbors import KNeighborsClassifier
import warnings
warnings.filterwarnings("ignore")
```

```
# Load dataset
df = pd.read_csv("/kaggle/input/exploring-mental-health-data/test.csv")

# Drop irrelevant columns
df = df.drop(columns=["id", "Name", "City", "Profession"])

# Separate numeric and categorical
num_cols = df.select_dtypes(include=['int64', 'float64']).columns
cat_cols = df.select_dtypes(include=['object']).columns

# Impute missing values
df[num_cols] = SimpleImputer(strategy="median").fit_transform(df[num_cols])
df[cat_cols] = SimpleImputer(strategy="most_frequent").fit_transform(df[cat_cols])

# Label encode categorical columns
le_dict = {}
for col in cat_cols:
    le = LabelEncoder()
    df[col] = le.fit_transform(df[col])
    le_dict[col] = le

# Features and target
X = df.drop(columns=["Working Professional or Student"])
y = df["Working Professional or Student"]

# Scale features
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)
```

```
# Define models
models = {
    "Logistic Regression": LogisticRegression(max_iter=1000),
    "Decision Tree": DecisionTreeClassifier(),
    "Random Forest": RandomForestClassifier(),
    "XGBoost": XGBClassifier(use_label_encoder=False, eval_metric='logloss')
}

# Plot ROC curves
plt.figure(figsize=(10, 8))

for name, model in models.items():
    model.fit(X_train, y_train)
    y_proba = model.predict_proba(X_test)[:, 1]
    fpr, tpr, _ = roc_curve(y_test, y_proba)
    roc_auc = auc(fpr, tpr)

    # Print classification report
    print(f"\n--- {name} ---")
    print("Classification Report:")
    print(classification_report(y_test, model.predict(X_test)))
    print("Confusion Matrix:")
    print(confusion_matrix(y_test, model.predict(X_test)))
```

```

--- Logistic Regression ---
Classification Report:
              precision    recall  f1-score   support

         0       0.61      0.55      0.58      3681
         1       0.89      0.91      0.90     15079

 accuracy      0.84      18760
 macro avg     0.75      18760
 weighted avg  0.84      18760

```

```

Confusion Matrix:
[[ 2016  1665]
 [ 1304 13775]]

```

```

--- Decision Tree ---
Classification Report:
              precision    recall  f1-score   support

         0       1.00      1.00      1.00      3681
         1       1.00      1.00      1.00     15079

 accuracy      1.00      18760
 macro avg     1.00      18760
 weighted avg  1.00      18760

```

```

--- Random Forest ---
Classification Report:
              precision    recall  f1-score   support

         0       1.00      1.00      1.00     3681
         1       1.00      1.00      1.00    15079

 accuracy          1.00      1.00      1.00    18760
 macro avg          1.00      1.00      1.00    18760
weighted avg          1.00      1.00      1.00    18760

Confusion Matrix:
[[ 3678    3]
 [    2 15077]]

--- XGBoost ---
Classification Report:
              precision    recall  f1-score   support

         0       1.00      1.00      1.00     3681
         1       1.00      1.00      1.00    15079

 accuracy          1.00      1.00      1.00    18760
 macro avg          1.00      1.00      1.00    18760
weighted avg          1.00      1.00      1.00    18760

```

39

Summary

1. Prescriptive analytics.
2. Big data.
3. Case study of data mining



Thank You