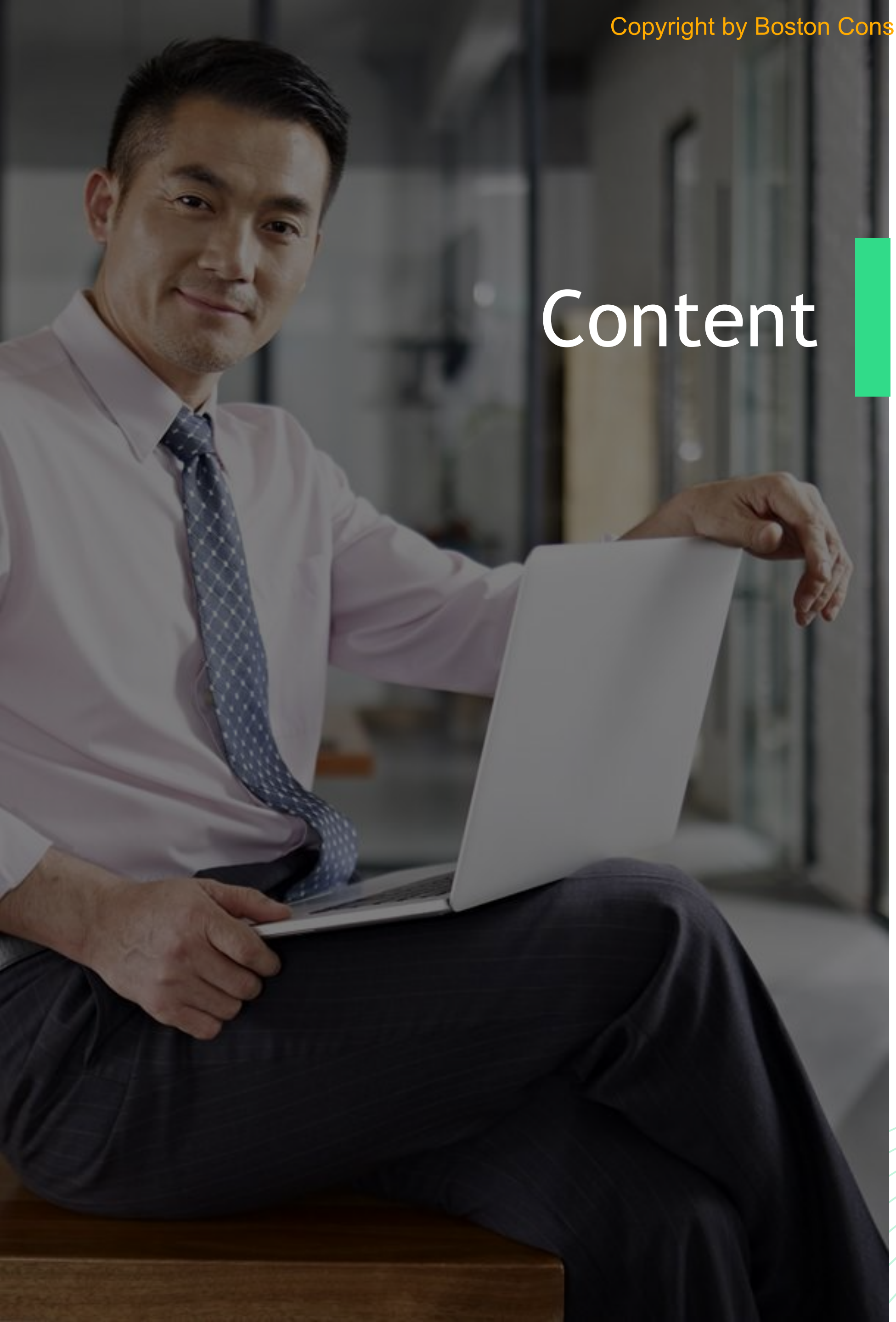
RISE BY DIGITALBCG ACADEMY

IU 4.6 – Machine Learning

Mini Project

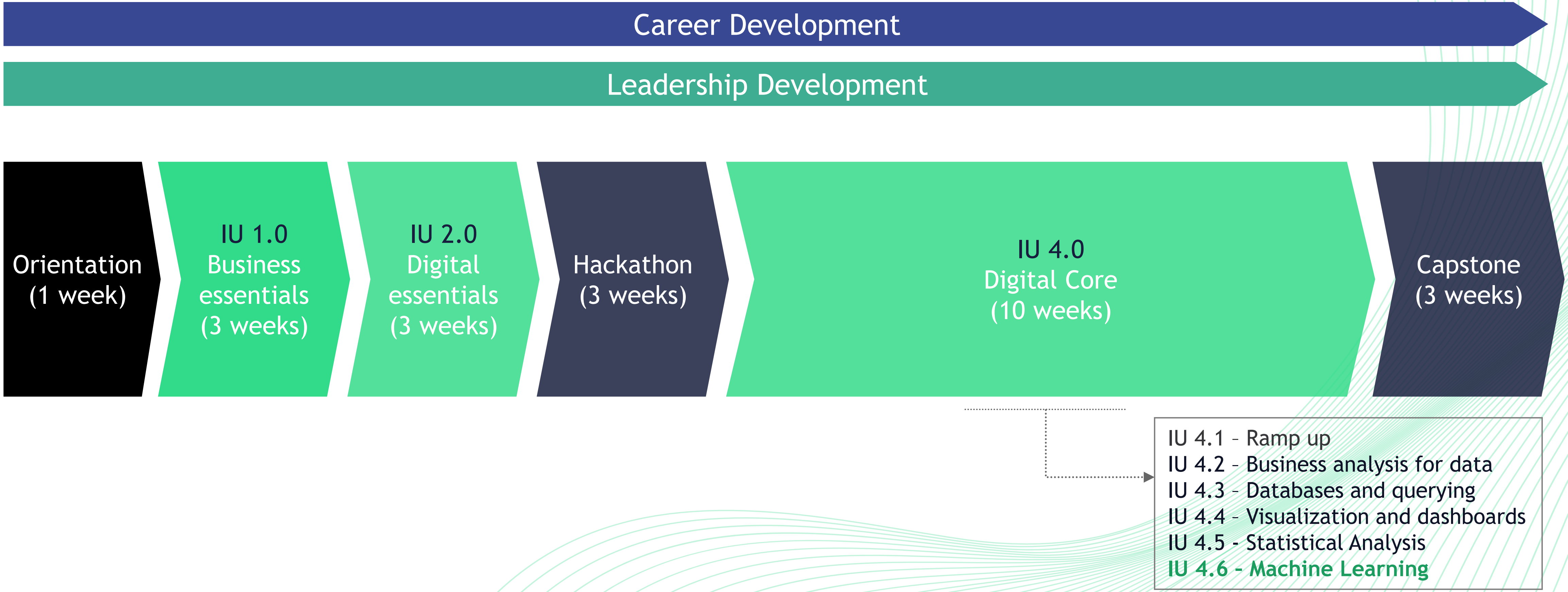
2021



Content

	180 min
Where Are We in the Journey & Learning Objectives	5 min
Agenda for Today	
• Mini Project: Introduction to problem statement	30 min
• Q&A	20 min
• Group activity to solve the project	120 min
Wrap up	5 min

Where are we in the learning journey?



Career Development

Leadership Development

Orientation
(1 week)

IU 1.0
Business essentials
(3 weeks)

IU 2.0
Digital essentials
(3 weeks)

Hackathon
(3 weeks)

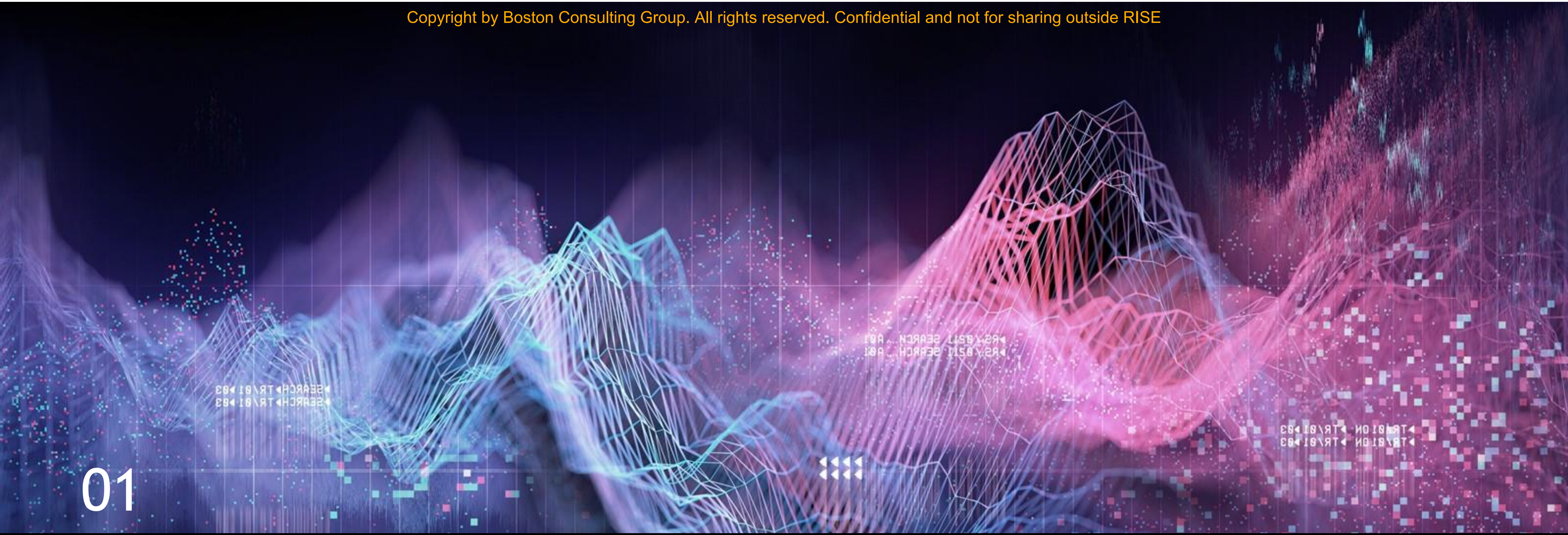
IU 4.0
Digital Core
(10 weeks)

Capstone
(3 weeks)

IU 4.1 – Ramp up
IU 4.2 – Business analysis for data
IU 4.3 – Databases and querying
IU 4.4 – Visualization and dashboards
IU 4.5 - Statistical Analysis
IU 4.6 - Machine Learning

Learning objectives

- Upon successful completion of the mini project, individuals will be able to build and evaluate machine learning models for a classification based problem
- Will be able to apply various model evaluation techniques from technical and business perspective



01

Mini Project: Introduction to the problem statement

Major SEA Telecom Provider

Providing telecom services to prepaid and postpaid customer segments combined with variety of product offers and plans

Business Problem

Company's postpaid business of voice only plans is struggling to maintain its strong foothold in local market because of,

- High churn rate amongst customers leading to a revenue decline of ~500k USD every month

- Decline in overall customer base (high churn rate combined with low acquisition rate), leading to a decline in total market share

Project objectives

Business Objective: Reduce monthly customer churn by identifying high risk customers well in advance

Hypothesis

- Company CEO believes that existing models can predict churners precisely, but it's too late to take any retention actions, as customers usage have significantly declined by then

Analytics Objective

1. Build a classification model to predict churners one month in advance

2. Identify key churn drivers

What is provided?

Data provided for 50k customers who are currently availing voice only postpaid plans from the telecom provider

Data Source	Data fields available
Customer profile	Months in service, Unique subscribers, Active subs, Service area, handset, model, Age HH1, Age HH2, Marital status, Occupation, Home ownership, Has Credit card, Owns motorcycle, Credit rating and changes over time, etc.
Customer Revenue	Monthly revenue, recurring revenue from prior month, Percentage change in revenue
Customer Usage and Activity	Monthly minutes, percent change in minutes, Overage minutes, Inbound & outbound calls, Peak & off-peak calls, dropped/blocked/unanswered calls, roaming calls, etc.
Customer Interaction	Customer care calls, Responds to mail offers, Retention calls, Retention offer accepted, referrals made, etc.
Churn	Churn/ Non-churn flag for each customer

Data Dictionary

Columns	Description
CustomerID	Unique Customer ID
MonthlyRevenue	Monthly revenue in USD
MonthlyMinutes	Monthly minutes
TotalRecurringCharge	Recurring charges in the past month
DirectorAssistedCalls	Automated calll (directory assisted calls)
OverageMinutes	Extra minutes above the postpaid allocation
RoamingCalls	Minutes on calls while roaming
PercChangeMinutes	percentage change in minutes from previous month
PercChangeRevenues	percentage change in revenue from previous month
DroppedCalls	Dropped calls in minutes
BlockedCalls	Blocked calls in minutes
UnansweredCalls	Unanswered calls in minutes
CustomerCareCalls	Customer care call duration in minutes
ThreewayCalls	Minutes spend on Conference calls
ReceivedCalls	Total received calls in minutes
OutboundCalls	Marketing calls received from customer service in minutes
InboundCalls	total duration in minutes of calls made to customer service
PeakCallsInOut	Incoming/outgoing calls during peak time
OffPeakCallsInOut	Incoming/outgoing calls during off peak time
DroppedBlockedCalls	Summation of dropped and blocked
CallForwardingCalls	Minutes in call forwarding
CallWaitingCalls	Minutes spend on hold during call
MonthsInService	Total months in service
UniqueSubs	total number of unique sim card subscriptions (includes inactive connection)
ActiveSubs	total # of active subscriptions
ServiceArea	Service area of the subscription
Handsets	Total # of handsets
HandsetModels	total # of unique hadnset model
CurrentEquipmentDays	Number of days since the activation of current equipment
AgeHH1	Primary holder
AgeHH2	Secondary holder
ChildrenInHH	Flag indicating childen in household
HandsetRefurbished	Handset refurbished flag (returned to company and then they sell it to different customer)
HandsetWebCapable	Internet connectivity
TruckOwner	Flag indicating whether the Customer owns a truck
RVOwner	Customer owns RV or not
Homeownership	Home owned by customer
BuysViaMailOrder	Whether customer has bought anything via clicking an option on email
RespondsToMailOffers	Flag indicating whether customer responds to mail offers
OptOutMailings	Whether the customer has opted out of mailing
NonUSTravel	Flag indicating whether the customer has travelled outside of US
OwnsComputer	whether customer Owns computers
HasCreditCard	Whether customer owns a credit card
RetentionCalls	Whether customer responded to retention calls
RetentionOffersAccepted	Whether customer accepted retention offers
NewCellphoneUser	New cellphone user flag
NotNewCellphoneUser	Total not new cellphone user flag
ReferralsMadeBySubscriber	Total referrals made by subscriber
IncomeGroup	Income group
OwnsMotorcycle	Owns motorcycle flag
AdjustmentsToCreditRating	Number of time the credit rating has ranged in past 1 year
HandsetPrice	Price of the handset in USD
MadeCallToRetentionTeam	Flag indicating whether customer made call to retention team
CreditRating	Credit rating of cthe customer
PrizmCode	Area group of customer home location
Occupation	Type of occupation
MaritalStatus	Marital status

Group activity | Assignment

Exercise

Work plan

Estimated time to complete: 5-7 hours

1. Detect and resolve problems in the data (Missing value, Outliers, Unexpected value, etc.) *[30 - 45 mins] [Marks: 10]*
   i. How many customers had zero monthly revenue?
   ii. How many columns have missing values percentage > 5%?
   iii. For columns, "UniqueSubs" and "DirectorAssistedCalls" remove outliers, if any

2. Perform exploratory analysis to analyze customer churn *[30 – 45 mins] [Marks: 15]*
   i. Does customers with high overage minutes also have high revenue?
   ii. Does high number of active subscribers lead to low monthly revenue?
   iii. Does credit rating have an impact in churn rate?

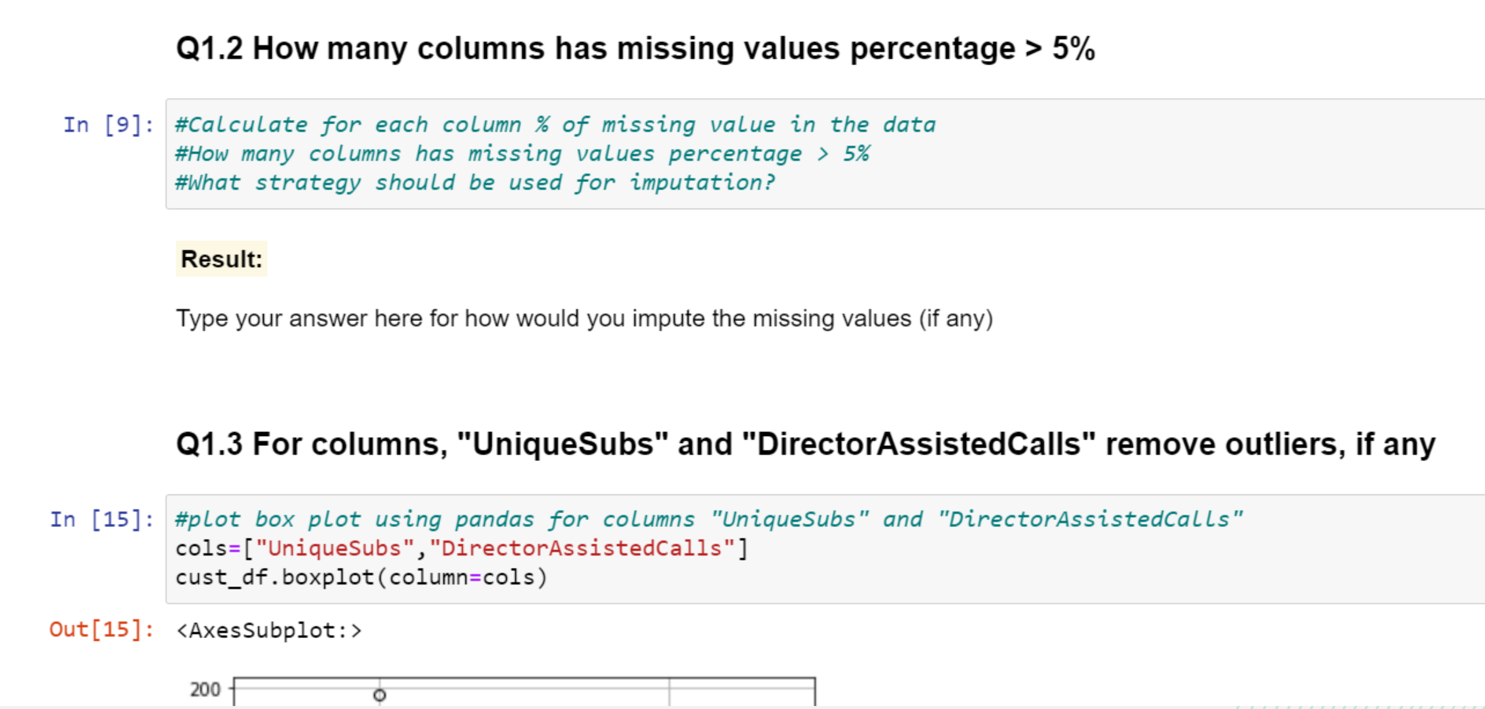
3. Create additional features to help predict churn *[20 – 40 mins] [Marks: 15]*
   i. Percent of current active subs over total subs
   ii. Percent of recurrent charge to monthly charge
   iii. Percent of overage minutes over total monthly minutes

4. Build classification model to predict customer churn *[120 - 180 mins] [Marks: 50]*
   i. Build a simple logistic regression model to predict churn and evaluate model accuracy on test data set
   ii. Build Random Forest classifier to compare model accuracy over the logistic regression model
   iii. Identify most important features impacting churn
   (Model evaluation metrics to be used: GINI, AUC, Precision and Recall)

5. Use the hold out data provided to predict churners using the best model identified in step 4 *[45 -60 mins] [Marks: 10]*

6. **Bonus Question:** Calculate lift chart and total monthly revenue saved by targeting top 10-20% of the customers using your best predictive model *[60 - 90 mins]*

What are we giving you?



Q1.2 How many columns has missing values percentage > 5%

```
In [9]: #Calculate for each column % of missing value in the data
        #How many columns has missing values percentage > 5%
        #What strategy should be used for imputation?
```

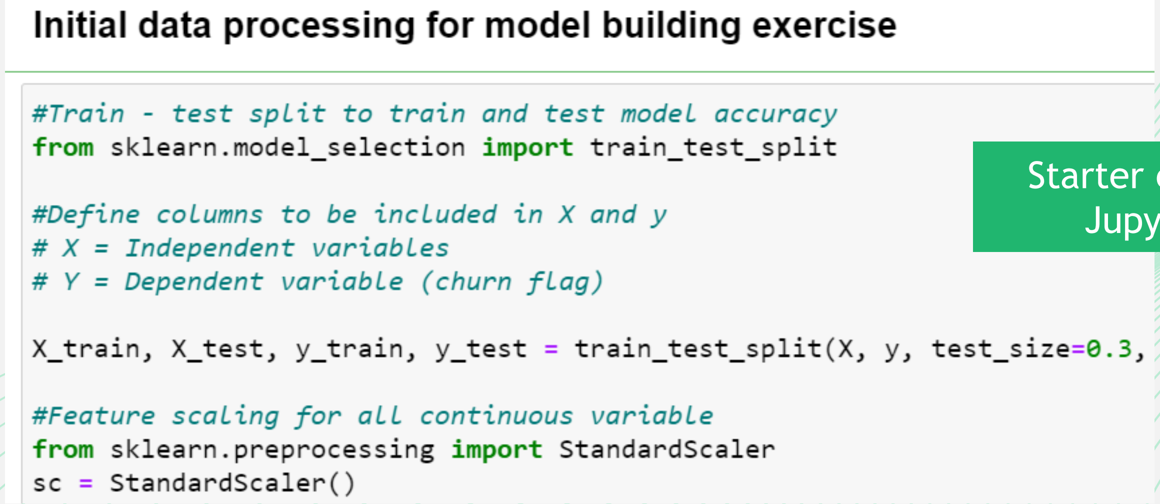
Result:

Type your answer here for how would you impute the missing values (if any)

Q1.3 For columns, "UniqueSubs" and "DirectorAssistedCalls" remove outliers, if any

```
In [15]: #plot box plot using pandas for columns "UniqueSubs" and "DirectorAssistedCalls"
         cols=["UniqueSubs","DirectorAssistedCalls"]
         cust_df.boxplot(column=cols)

Out[15]: <AxesSubplot:>
```



Initial data processing for model building exercise

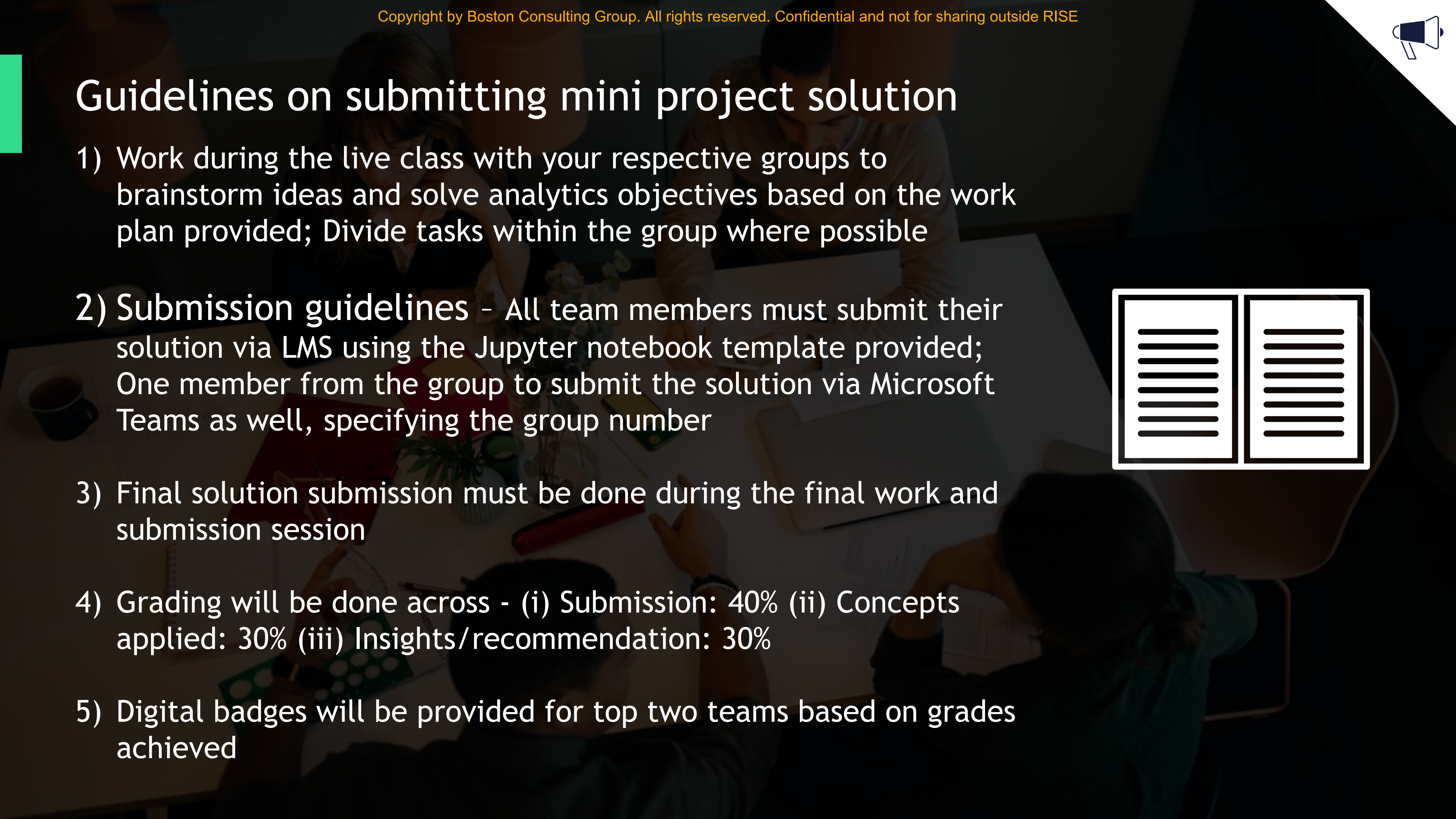
```
#Train - test split to train and test model accuracy
from sklearn.model_selection import train_test_split

#Define columns to be included in X and y
# X = Independent variables
# Y = Dependent variable (churn flag)

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3,

#Feature scaling for all continuous variable
from sklearn.preprocessing import StandardScaler
sc = StandardScaler()
```

Starter code provided in Jupyter notebook



Guidelines on submitting mini project solution

1) Work during the live class with your respective groups to brainstorm ideas and solve analytics objectives based on the work plan provided; Divide tasks within the group where possible

2) Submission guidelines – All team members must submit their solution via LMS using the Jupyter notebook template provided; One member from the group to submit the solution via Microsoft Teams as well, specifying the group number

3) Final solution submission must be done during the final work and submission session

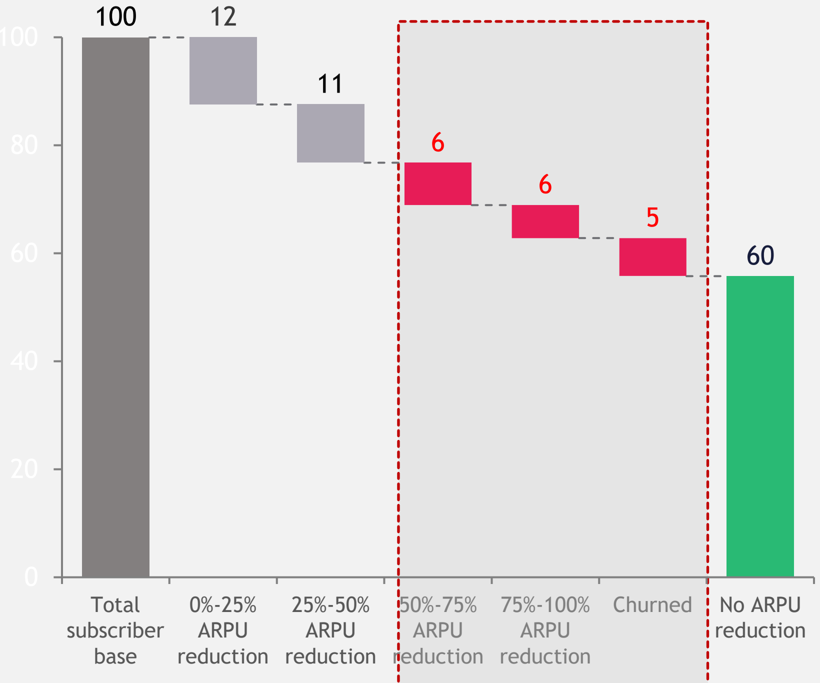
4) Grading will be done across - (i) Submission: 40% (ii) Concepts applied: 30% (iii) Insights/recommendation: 30%

5) Digital badges will be provided for top two teams based on grades achieved

Recap Customer Churn - Typical business scenarios in Telco industry

Steady decline in ARPU

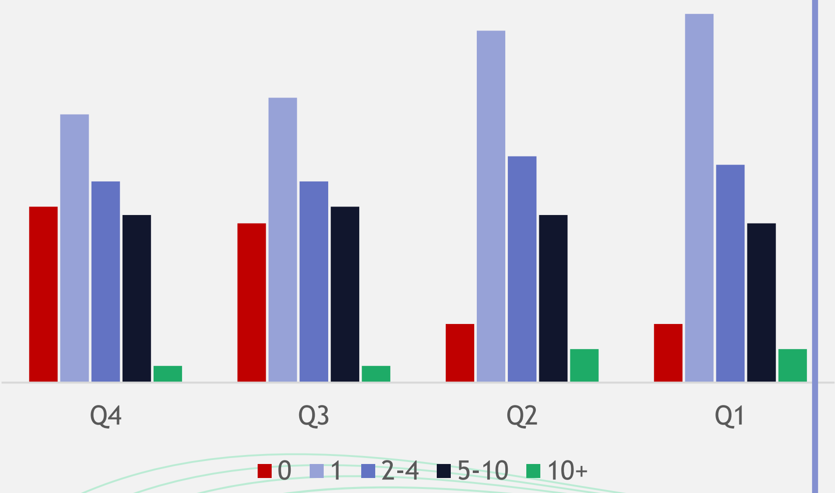
About **X%** customers exhibit more than 50% drop in monthly ARPU for 3 months straight



Group of subscribers by ARPU changes

Low Customer Engagement

It was observed that the average number of recharges per customer had gone down by **Y%** in last 6 months



QoQ trend of number of recharges per customer per month

Lack of Retention Strategy

The retention offers sent out to customers are driven by a basic rule-based framework

Key drivers for existing retention offer framework

- Last Recharge Amount
- Current Plan
- Best available discount

Retention offers sent as a part of the same campaigns as up-sell/growth

Recap: Machine Learning topics that you will need to use

1 Exploratory data analysis

- Missing value identification and treatment, outlier detection, etc.
- Univariate analysis - histograms to check distribution, box/violin plot, summary stats such as mean, median, mode, etc.
- Correlation analysis, scatter plots
- Time analysis to monitor trend and seasonality

2 Feature engineering

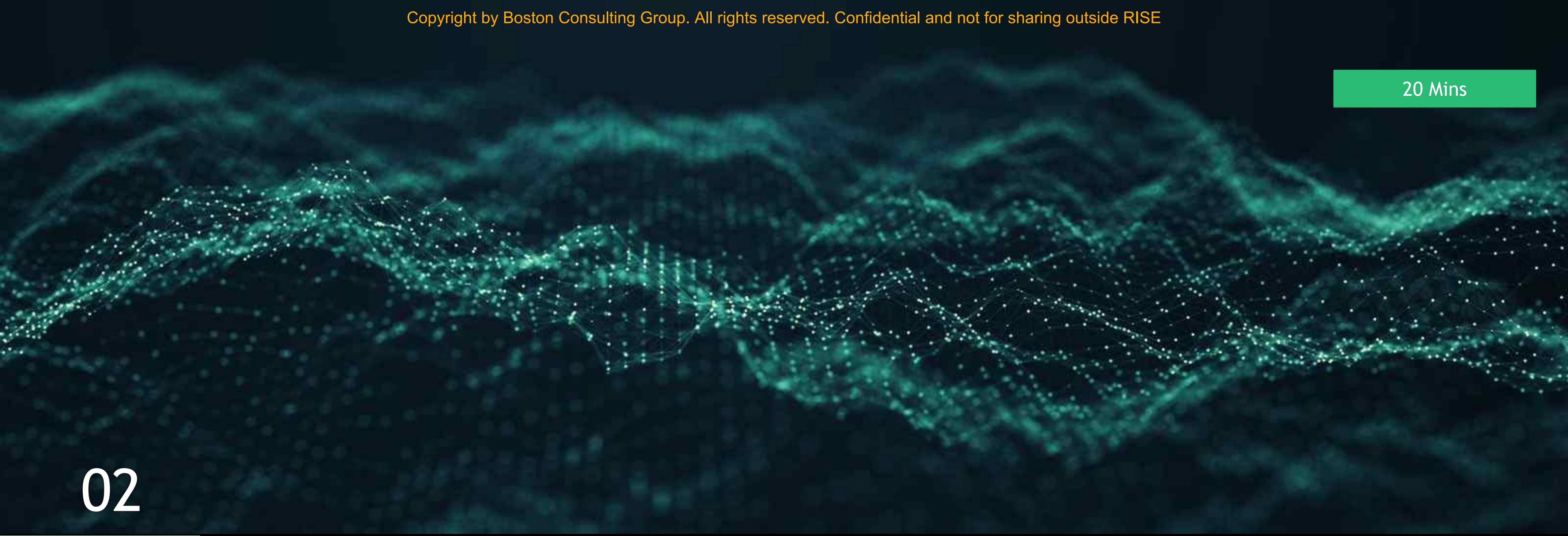
- Target variable creation (dependent variable) - align with analytics objective
- New feature creation based on business and analytics objectives
- Trend variables such as moving average, standard deviation over time, etc.

3 ML model build

- Model type – classification vs regression
- Algorithm type – logistic regression, linear regression, Random forest, etc.
- Train and test data split (e.g., 70:30 split)
- Hyperparameter tuning for model optimization
- Feature importance

4 Model evaluation

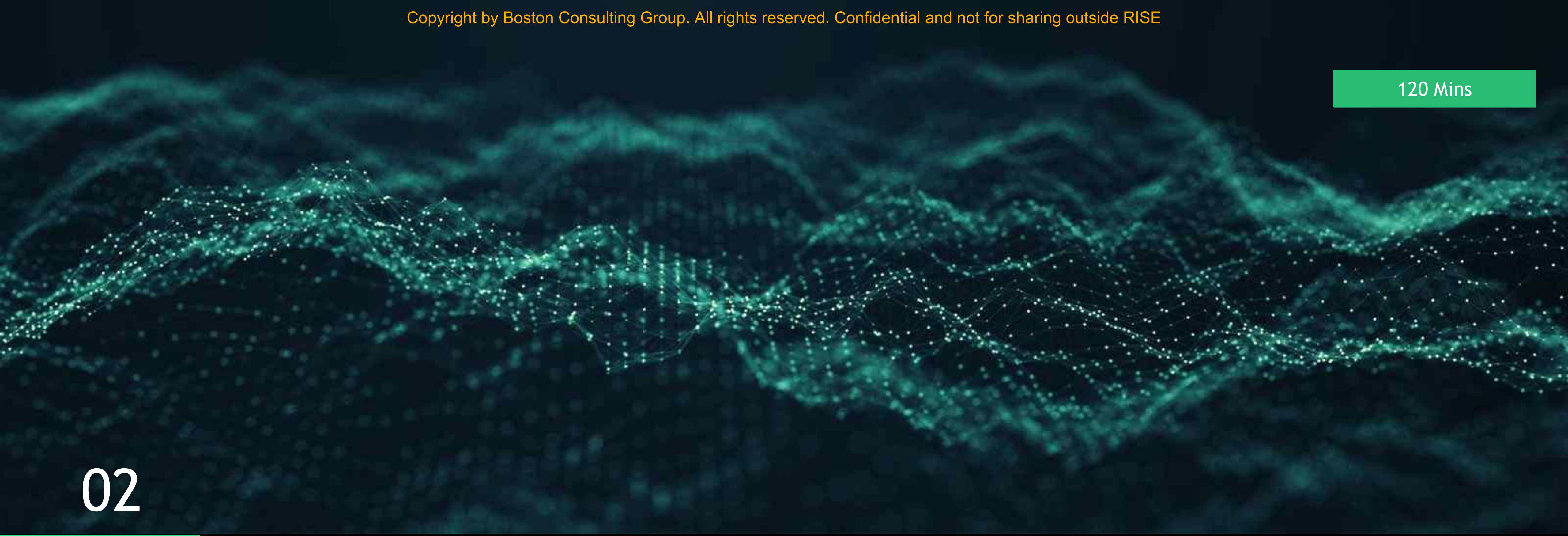
- Model evaluation metrics
- Regression – R-Square, Mean Absolute Error (MAE), Mean Square Error (MSE), etc.
- Classification – Confusion metrics, precision, recall, GINI, AUC, etc.
- Lift and Gain charts
- Hold out data set for model validation



20 Mins

02

Questions?



120 Mins

02

Working session - Group activity

RISE BY DIGITAL BCG ACADEMY