W05 – W08
SPARK MACHINE LEARNING

2024

The Hoang

LEARNING OBJECTIVES

Learning the Apache Spark ML by Practicing Labs:

- Customer Churn Analysis

APACHE SPARK'S MLLIB – SUMMARY

Recap

Category	Algorithms in Spark MLlib
Supervised Learning	Logistic Regression, Decision Trees, Random Forests, GBT, Linear Regression, Naive Bayes, etc.
Classification	Logistic Regression, Decision Trees, Random Forests, GBT, SVM, Naive Bayes, MLP
Regression	Linear Regression, Decision Trees, Random Forests, GBT, Generalized Linear Models, AFT
Unsupervised Learning	K-means, GMM, Bisecting K-means, LDA, PCA, SVD
Clustering	K-means, GMM, Bisecting K-means, LDA
Dimensionality Reduction	PCA, SVD
Recommendation Systems	ALS (Alternating Least Squares)
Ensemble Methods	Random Forests, Gradient-Boosted Trees
Feature Transformation	StandardScaler, MinMaxScaler, StringIndexer, OneHotEncoder, TF-IDF, Word2Vec, etc.
Evaluation & Tuning	CrossValidator, TrainValidationSplit, BinaryClassificationEvaluator, RegressionEvaluator, etc.

DATA PREPROCESSING & CLEANSING

Importance of Data Preprocessing and Cleansing

- **Quality of Insights**: Clean data leads to more accurate and reliable insights. Poor quality data can skew results and lead to incorrect conclusions.
- **Model Performance**: Machine learning models are sensitive to the quality of the input data. Preprocessing helps improve the model's performance by ensuring that the data is in a suitable format.
- **Reducing Noise**: Data often contains noise (irrelevant or redundant information). Cleaning the data helps to reduce this noise, allowing models to focus on relevant patterns.
- **Handling Missing Values**: Real-world data is often incomplete. Proper handling of missing values prevents biases and inaccuracies in analysis.
- **Feature Engineering**: Preprocessing allows data scientists to create new features that can improve model effectiveness.

DATA PREPROCESSING & CLEANSING

Typical Activities in Data Preprocessing and Cleansing

- **Data Cleaning:**
  - Removing Duplicates: Identifying and eliminating duplicate records to ensure data integrity.
  - Handling Missing Values: Strategies include imputation, removal, or using algorithms that can handle missing values.
  - Filtering Outliers: Identifying and addressing outliers that may skew results.
- **Data Transformation:**
  - Normalization/Standardization: Scaling numerical features to a common range to improve model training.
  - Encoding Categorical Variables: Converting categorical data into numerical format using techniques like one-hot encoding or label encoding.
- **Data Reduction:**
  - Dimensionality Reduction: Techniques like PCA (Principal Component Analysis) to reduce the number of features while retaining essential information.
  - Sampling: Reducing the size of the dataset by selecting a representative subset.

DATA PREPROCESSING & CLEANSING

Typical Activities in Data Preprocessing and Cleansing

- **Data Integration:**
  - Merging Datasets: Combining data from different sources to create a unified dataset for analysis.
  - Resolving Schema Conflicts: Ensuring that data from different sources aligns properly and follows a consistent schema.
- **Data Formatting:**
  - Date and Time Formatting: Ensuring date and time fields are in a consistent format.
  - String Manipulation: Cleaning and formatting string data (e.g., trimming whitespace, correcting typos).
- **Feature Engineering:**
  - Creating New Features: Deriving new variables that can provide additional insights or improve model accuracy.
  - Binning: Converting continuous variables into categorical ones by grouping them into bins.

DATA PREPROCESSING & CLEANSING

LAB: Preprocessing & Cleansing Data

CUSTOMER CHURN ANALYSIS

Customer churn is defined as when customers or subscribers discontinue doing business with a firm or service.

Purpose:

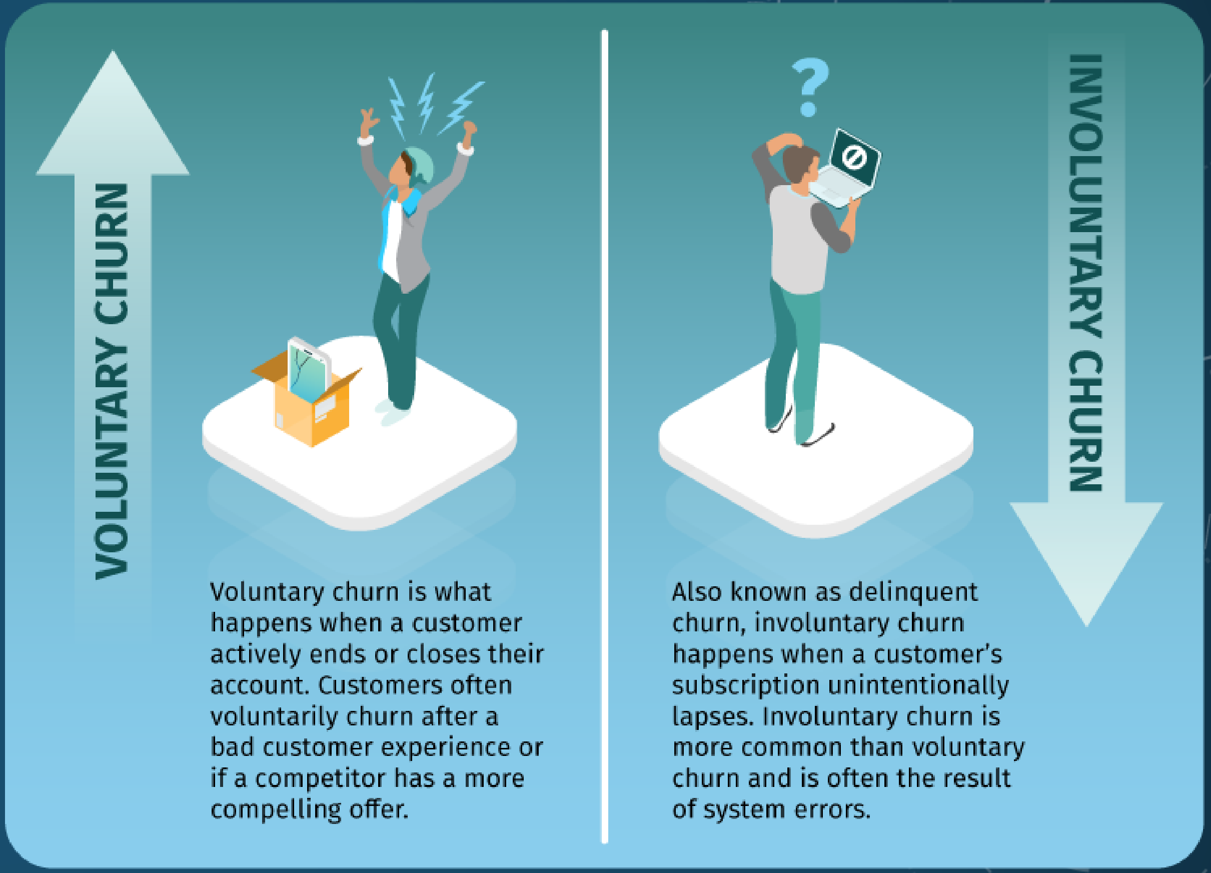
- Forecast / predict which customers are likely to leave ahead of time.
- Focus customer retention efforts only on these "high risk" clients.
- The ultimate goal is to expand its coverage area and retrieve more customers loyalty**.**
- Customer churn is a critical metric because it is much less expensive to retain existing customers than it is to acquire new customers

CUSTOMER CHURN ANALYSIS

To detect early signs of potential churn, one must first develop a holistic view of the customers and their interactions across numerous channels, including store/branch visits, product purchase histories, customer service calls, Web-based transactions, and social media interactions …

As a result, by addressing churn, these businesses may not only preserve their market position but also grow and thrive.

More customers they have in their network, the lower cost of initiation and the larger the profit. As a result, the company's key focus for success is reducing client attrition and implementing effective retention strategy.



Voluntary churn is what happens when a customer actively ends or closes their account. Customers often voluntarily churn after a bad customer experience or if a competitor has a more compelling offer.

Also known as delinquent churn, involuntary churn happens when a customer's subscription unintentionally lapses. Involuntary churn is more common than voluntary churn and is often the result of system errors.

CUSTOMER CHURN ANALYSIS

LAB: Customer churn analysis

CUSTOMER SEGMENTATION

Customer segmentation is a technique used to divide customers into distinct groups based on their behaviour, demographics, preferences, and other characteristics.

Objectives: The objectives of customer segmentation are to:

- Identify high-value customers from a large customer base
- Understand customer behaviour and preferences
- Develop targeted marketing campaigns
- Improve customer satisfaction and loyalty

Case studies:

A telecom company used customer segmentation to identify customers who were likely to churn and developed targeted retention campaigns, resulting in a 30% reduction in churn.

CUSTOMER SEGMENTATION

Required datasets:

- Customer demographic data (e.g. age, gender, location)
- Customer behaviour data (e.g. purchase history, browsing history)
- Customer preference data (e.g. product preferences, communication preferences)

Steps:

1. Collect and clean the data
2. Identify the variables to use for segmentation (e.g. demographic, behavioral, preference)
3. Choose a clustering algorithm (e.g. K-Means, Hierarchical Clustering)
4. Apply the clustering algorithm to the data
5. Evaluate the clusters and identify the most valuable segments
6. Develop targeted marketing campaigns for each segment

CUSTOMER SEGMENTATION

LAB: Customer Segmentation

PROPENSITY MODELLING

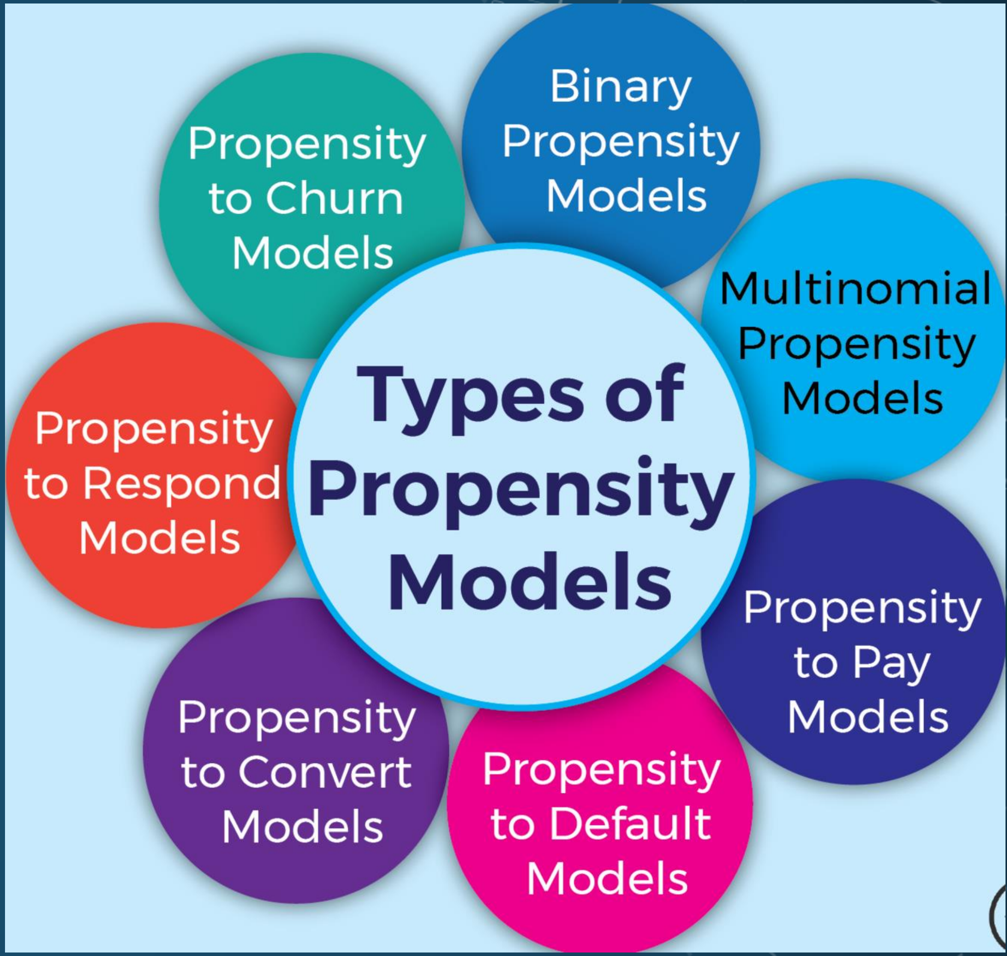
Propensity modeling is a technique used to predict the likelihood of a customer to perform a specific action (e.g. make a purchase, respond to an offer, customer lifetime value, propensity to churn).

Objectives: The objectives of propensity model are to:

- Identify customers who are likely to perform a specific action
- Predict customer behaviour
- Develop targeted marketing campaigns
- Improve customer conversion rates

Case studies:

A financial services company used propensity modeling to predict which customers were likely to respond to an offer and developed targeted marketing campaigns, resulting in a 30% increase in response rates.



Types of Propensity Models

Propensity to Churn Models

Binary Propensity Models

Multinomial Propensity Models

Propensity to Respond Models

Propensity to Pay Models

Propensity to Convert Models

Propensity to Default Models

PROPENSITY MODELLING

Required datasets:

- Customer demographic data (e.g. age, gender, location)
- Customer behaviour data (e.g. purchase history, browsing history)
- Customer preference data (e.g. product preferences, communication preferences)
- Historical data on customer actions (e.g. purchases, responses to offers)

Steps:

1. Collect and clean the data
2. Identify the variables to use for segmentation (e.g. demographic, behavioural, preference)
3. Choose a modeling algorithm (e.g. Logistic regression, Decision trees, …)
4. Apply the modeling algorithm to the data
5. Evaluate the model and identify the most predictive variables
6. Use the model to predict customer propensity and develop targeted marketing campaigns.

PROPENSITY MODELLING

LAB: Propensity Modeling

NEXT BEST ACTION (NBA)

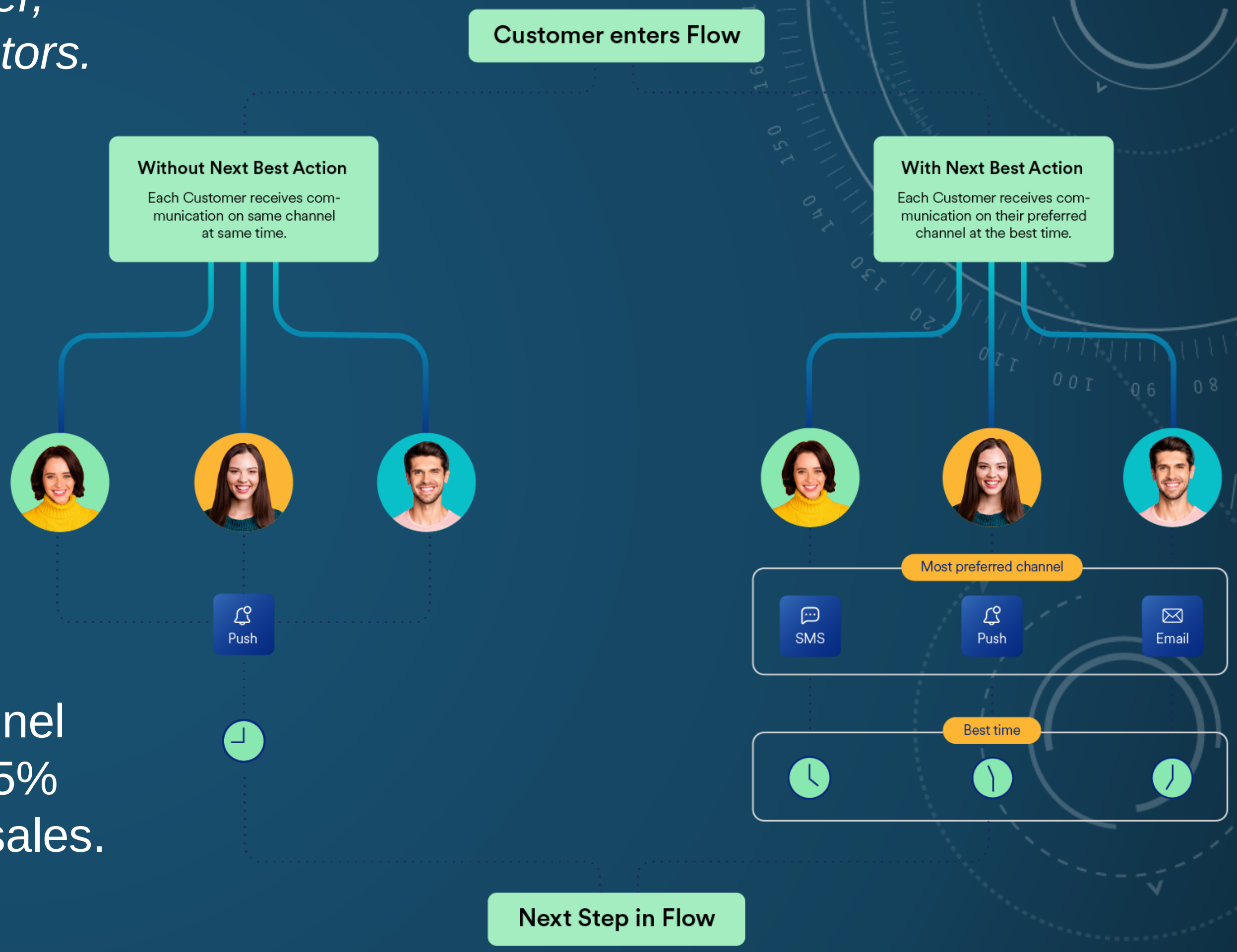
Definition: *Next Best Action (NBA) is a technique used to determine the most effective action to take with a customer, based on their past behaviour, preferences, and other factors.*

Objectives: The objectives of NBA are to:

- Improve customer satisfaction and loyalty
- Increase sales and revenue
- Enhance customer experience

Case studies:

Company used NBA to determine the most effective channel to send customers marketing information, resulting in a 25% increase in customer satisfaction and a 15% increase in sales.



Customer enters Flow

Without Next Best Action

Each Customer receives communication on same channel at same time.

With Next Best Action

Each Customer receives communication on their preferred channel at the best time.

Next Step in Flow

NEXT BEST ACTION (NBA)

Required datasets:

- Customer demographic data (e.g. age, gender, location)
- Customer behavior data (e.g. purchase history, browsing history)
- Customer preference data (e.g. product preferences, communication preferences)
- Customer interaction data (e.g. website interactions, customer service interactions)

Steps:

1. Collect and clean the data
2. Identify the variables to use for NBA (e.g. demographic, behavioral, preference)
3. Choose a machine learning algorithm (e.g. decision trees, random forests)
4. Train the model using the data
5. Evaluate the model and identify the most effective actions
6. Implement the actions and monitor the results

NEXT BEST ACTION (NBA)

LAB: Next Best Action

OTHER POPULAR CUSTOMER ANALYSIS

- Next Best Offer (NBO): NBO is a technique used to determine the most relevant offer or recommendation for a customer based on their past behavior, preferences, and other factors.
- Customer Journey Mapping: This involves analyzing the customer's journey across multiple touchpoints and channels to identify pain points, opportunities, and areas for improvement.
- Sentiment Analysis: Sentiment analysis involves analyzing customer feedback, reviews, and social media posts to understand their emotions and opinions about a product or service.
- Predictive Maintenance: Predictive maintenance involves using machine learning algorithms to predict when a customer is likely to experience a problem or issue with a product or service.
- Personalization: Personalization involves using AI/ML to create tailored experiences for customers based on their behavior, preferences, and other characteristics.
- Customer Lifetime Value (CLV): CLV involves analyzing the total value of a customer over their lifetime, including their past purchases, future potential, and other factors.
- Customer Health Score: Customer health score involves analyzing various metrics, such as engagement, satisfaction, and loyalty, to determine the overall health of a customer relationship.