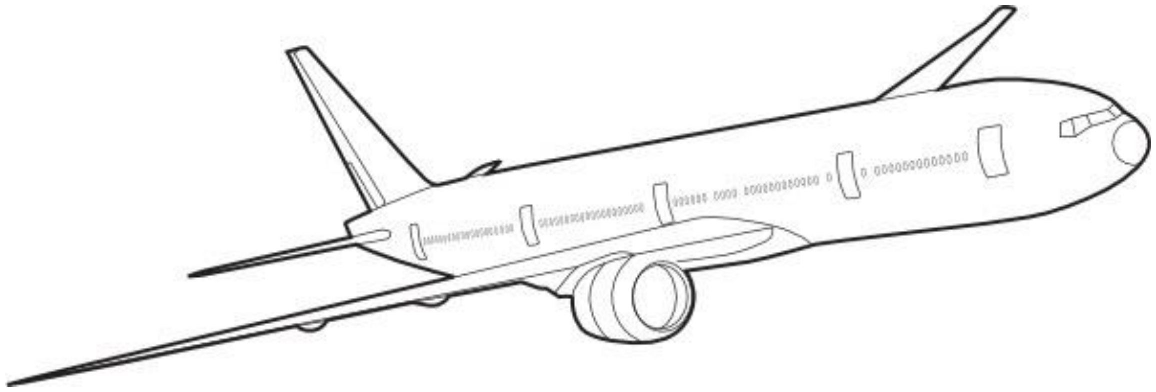
APACHE Spark™



A tutorial on Spark ML

&

Analyzing the customer patterns of ABC Airlines

December 05, 2017, A project By: Abhijit, Abhilasha, Ankit, Ao, and Suzanne

Agenda

- *Tutorial on Spark ML*
  - *K-means clustering (Unsupervised learning)*
  - *Topic Modeling using LDA (Unsupervised)*
  - *Logistic Regression (Supervised learning)*

Clustering

Unsupervised Learning

(K-means)

ABC Airlines would like to customize marketing by identifying actionable customer groups

Current State	*Desired State*	*Why Big Data*
• *ABC Airlines is a Minneapolis based regional airlines which flies customers to warmer destinations* • *ABC has a strict marketing budget and would like to optimize marketing by understanding various customer groups that exist*	• *ABC has marketing focused actionable groups of customers* • *ABC is able to maximize conversion based on the limited budget*	• *ABC Airlines has a huge and growing data stream that cannot be successfully processed on standalone machines* • *ABC is interested in kickstarting a distributed computing framework since it's more cost efficient*

Features available for clustering

Booking Channel

Base Fare

Age group

Seasonality

Originating from MSP

Pre-booking

people

Origin

Destination

If group ticket

Group size

Loyalty program membership

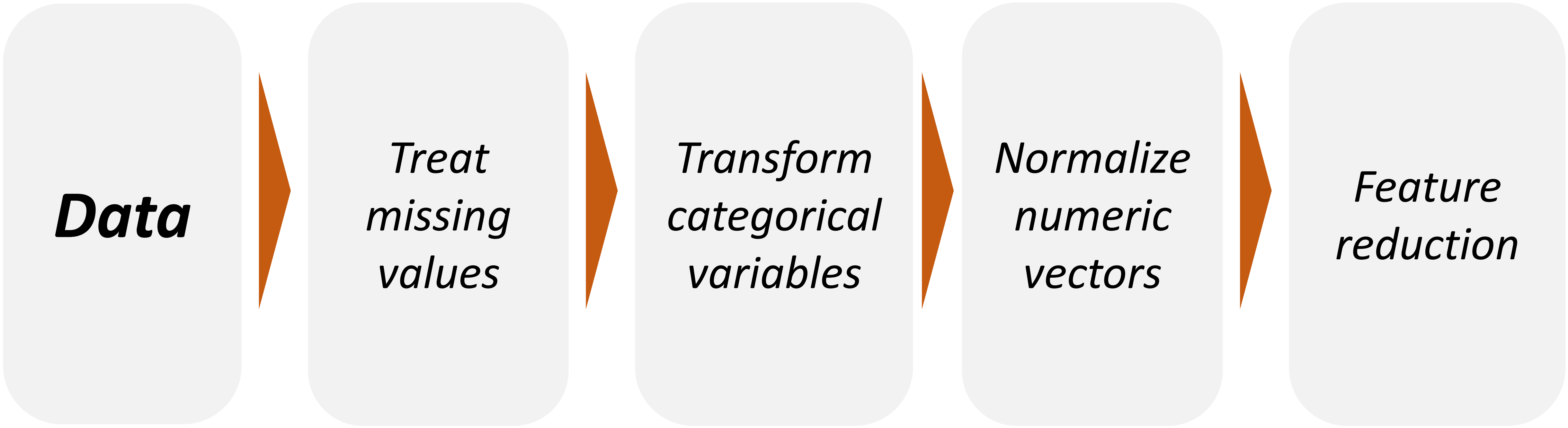
Loyalty program status

Flight date

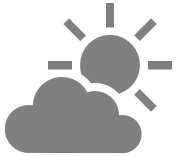
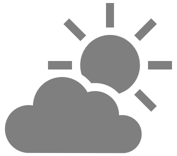
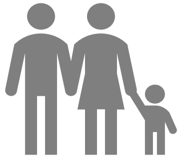
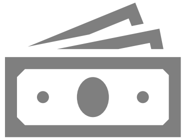
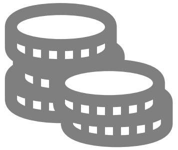
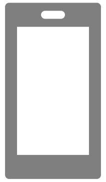
PNR

Passenger name

Data pre-processing



ABC has distinct segment of customers who can be targeted with customized strategy

	Vacationing Grandparents	Young Professionals
Booking Period		
Age		
Ticket Price		
Booking Channel		

LDA

Unsupervised Learning

(Topic modeling)

ABC Airlines would like to customize marketing by identifying actionable customer groups

Current State	***Desired State***	***Why Big Data***
• *ABC Airlines is a Minneapolis based regional airlines which flies customers to warmer destinations* • *ABC would like to understand the customers' sentiments from the reviews on third party websites*	• *ABC understand the major topics being discussed by customers on various travel websites* • *ABC is able to strategize to effectively mitigate major concerns*	• *ABC Airlines has a huge and growing data that cannot be successfully processed on standalone machines* • *ABC is interested to kickstart the distributed computing framework since it's more cost efficient*

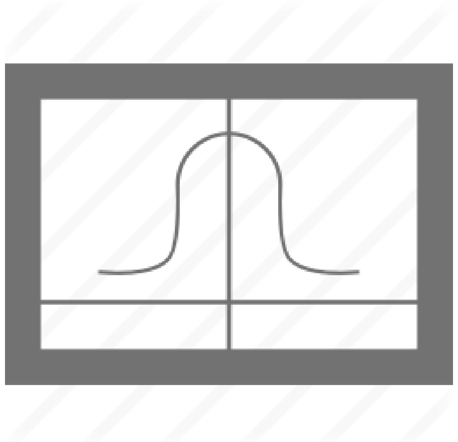
Data is cleaned and normalized before running LDA model

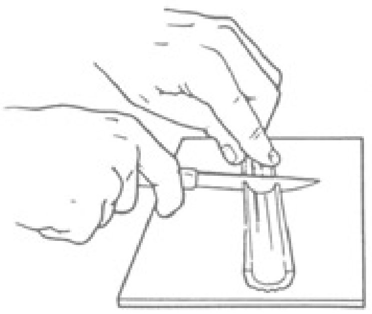
Normalize

Removed punctuations

Remove stop words

Stemming



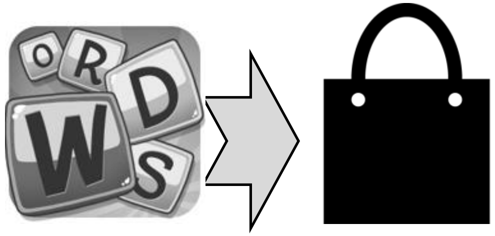


Latent Dirichlet Allocation (LDA) model was used to obtain topics from the customer reviews

Bag of Words

LDA

Top Topics



What the customers talk about on Social Media

Crew Appreciation

- Best, Crew, Think, Seemed

Vacation

- Vacation, Mexico, Took, Lots

On-time Performance

- Appreciate, Early, Sure, Plane

Random Forest

Supervised Learning

XYZ Corp would like to develop a custom spam filter for its employee emails

Current State	*Desired State*	*Why Big Data?*
• *XYZ is a International conglomerate operating in many domains and geographies* • *XYZ employees receive thousands of emails on a daily basis* • *As an improved security measure XYZ would like implement a custom spam filter for its employee emails*	• *XYZ has predictive algorithms to identify spam emails* • *XYZ has the agility to customize the spam filter for various departments*	• *XYZ Corp has a huge and growing email repository that cannot be successfully processed on standalone machines* • *XYZ is interested to kickstart the distributed computing framework since it's more cost efficient*

Most commonly used spam words were identified as features for the predictive model

Remove

Internet

order

mail

Will

Credit

Your

Font

people

000

Money

Free

Addresses

(

[

$

#

Capital Run Length Average

Capital Run Length Longest

Capital Run Length Total

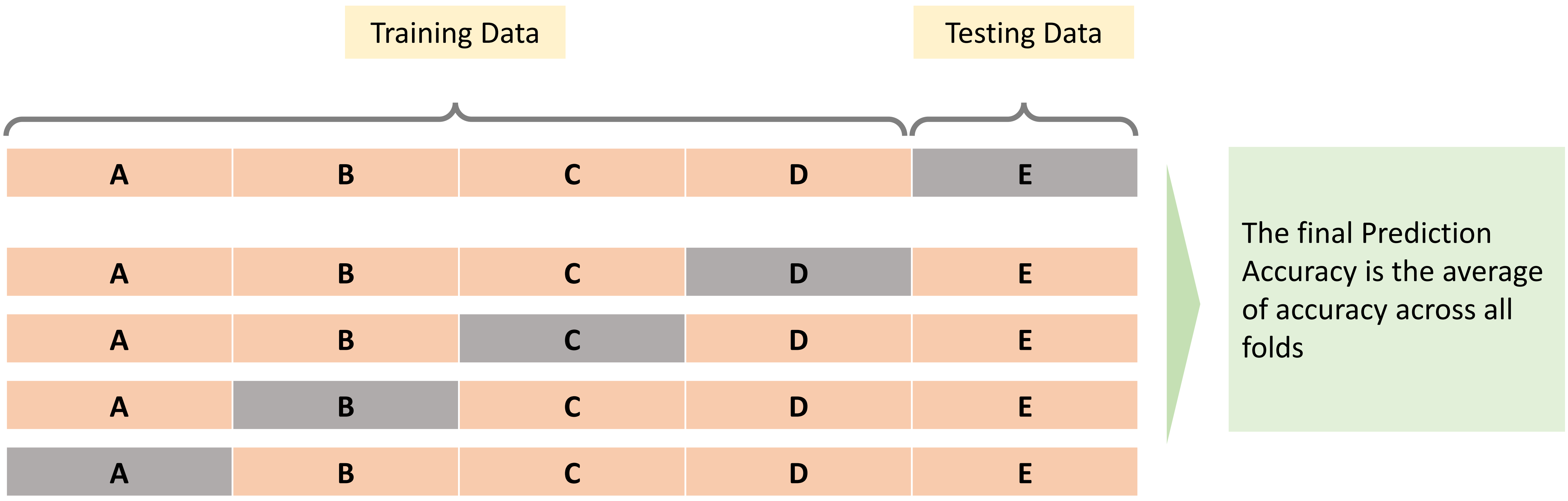
Data needs to be treated and transformed before passing it to classification models

Select Data → *Formatting* → *Cleaning* → *Sampling* → *Transform*

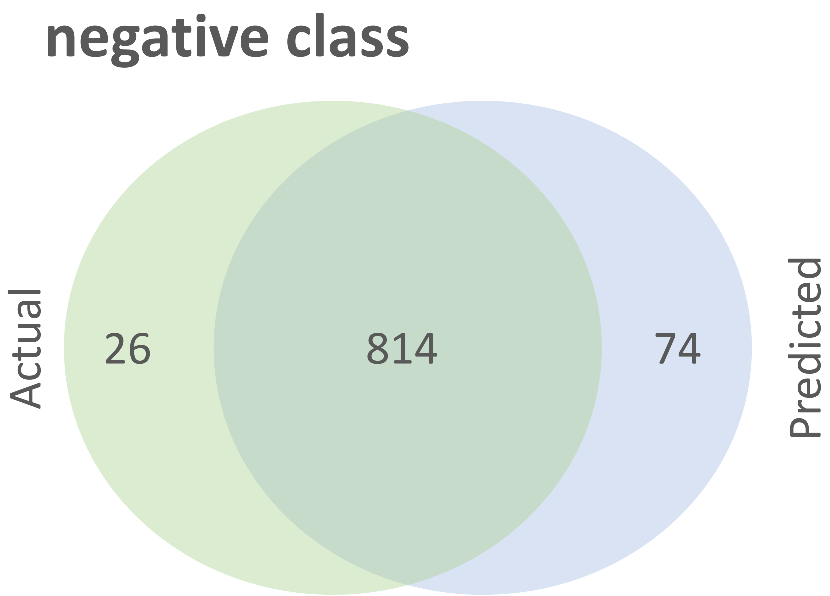
- *Scaling*
- *Decomposition*
- *Aggregation*

Cross validation allows users to check the robustness of the model

We used 5 fold cross validation to predict the spam emails.
This allows us to divide the 4,601 instances into 5 equal chunks.
The cross validation runs the model 5 times keeping one of the chunks as test and using the rest 4 for training



XYZ is able to identify spams using random forest with an accuracy of more 90%



Accuracy	92.7%
Recall	86.0%
Precision	94.6%

Qubole makes working on big data with teams easy

- **Easy collaboration - working with teams, GitHub**
  - Customize permission/access to team members
- **Cloud agnostic - Connects to:**
  - AWS, Google Cloud, Microsoft Azure
- **Support for multiple open source engines like:**
  - Hadoop, Hive, Apache Spark
- **Integrated Zeppelin Notebook:**
  - allow users to run language of their choice

Appendix

- *Clusters*
- *Datasets*
- *References*

Clusters › Cluster (41815)

1 master, 2 slave configuration was used for classification

Stop Edit

S 41815 test default

Status: Up

Nodes : 2 (0)

Up Time: 1h 41m

Cluster Type: Spark

Started at: 03 Dec 2017 10:14:43

Master DNS: ec2-52-87-239-50.compute-1.amazonaws.com

Resources

Resource Manager

DFS Status

AutoScaling Logs

Spark History Server

Cluster Start Logs

Nodes

Instance Type	Role	Public DNS	Private IP	Spot Instance	Up Time	Node Bootstrap Logs
r3.xlarge	Master	ec2-52-87-239-50.compute-1.amazonaws.com	ip-172-31-89-46.ec2.internal	No	1h 32m	View
r3.2xlarge	Slave	ec2-54-89-252-120.compute-1.amazonaws.com	ip-172-31-86-250.ec2.internal	No	1h 32m	View
r3.2xlarge	Slave	ec2-54-172-24-209.compute-1.amazonaws.com	ip-172-31-84-141.ec2.internal	No	1h 32m	View