MANIPAL INSTITUTE OF TECHNOLOGY
MANIPAL
(A constituent institution of MAHE, Manipal)

CUSTOMER SEGMENTATION
FINAL PRESENTATION

Atul Virendra Poddar
Mechanical Engineering – 170909178
Guide : Prof. Vinyas

AGENDA

- Quick Recap
- Correlation and VIF analysis
- Solution Design
- Algorithms Implemented
- Model Winner
- Business Impact

*Variable Inflation Factor

PROBLEM STATEMENT - RECAP

The credit card business (X) of a big conglomerate is interested in capitalizing untapped acquisition potential within its movie customer base (Y)

Problem at hand:
How to identify and acquire profitable customers for X from Y ?

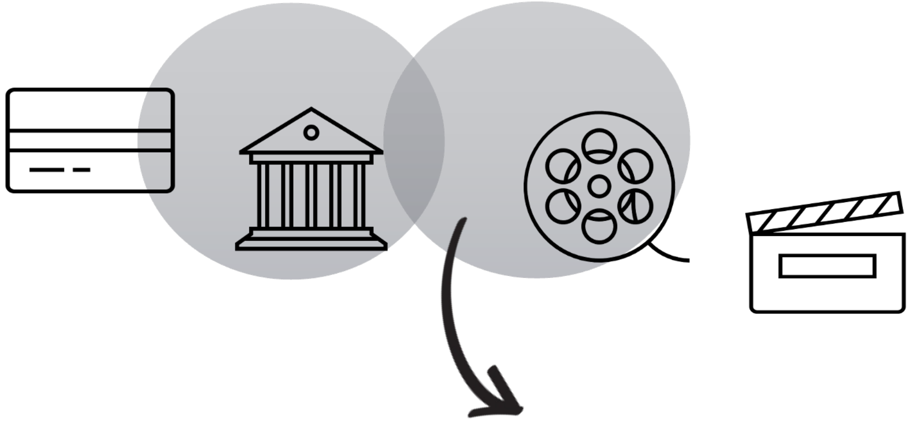
Analytics Problem

- To **understand** the **behaviour of customers** who use X credit cards for payments at VOX cinemas
- To **identify profitable customers** who will purchase X credit cards

Analytics Outcome

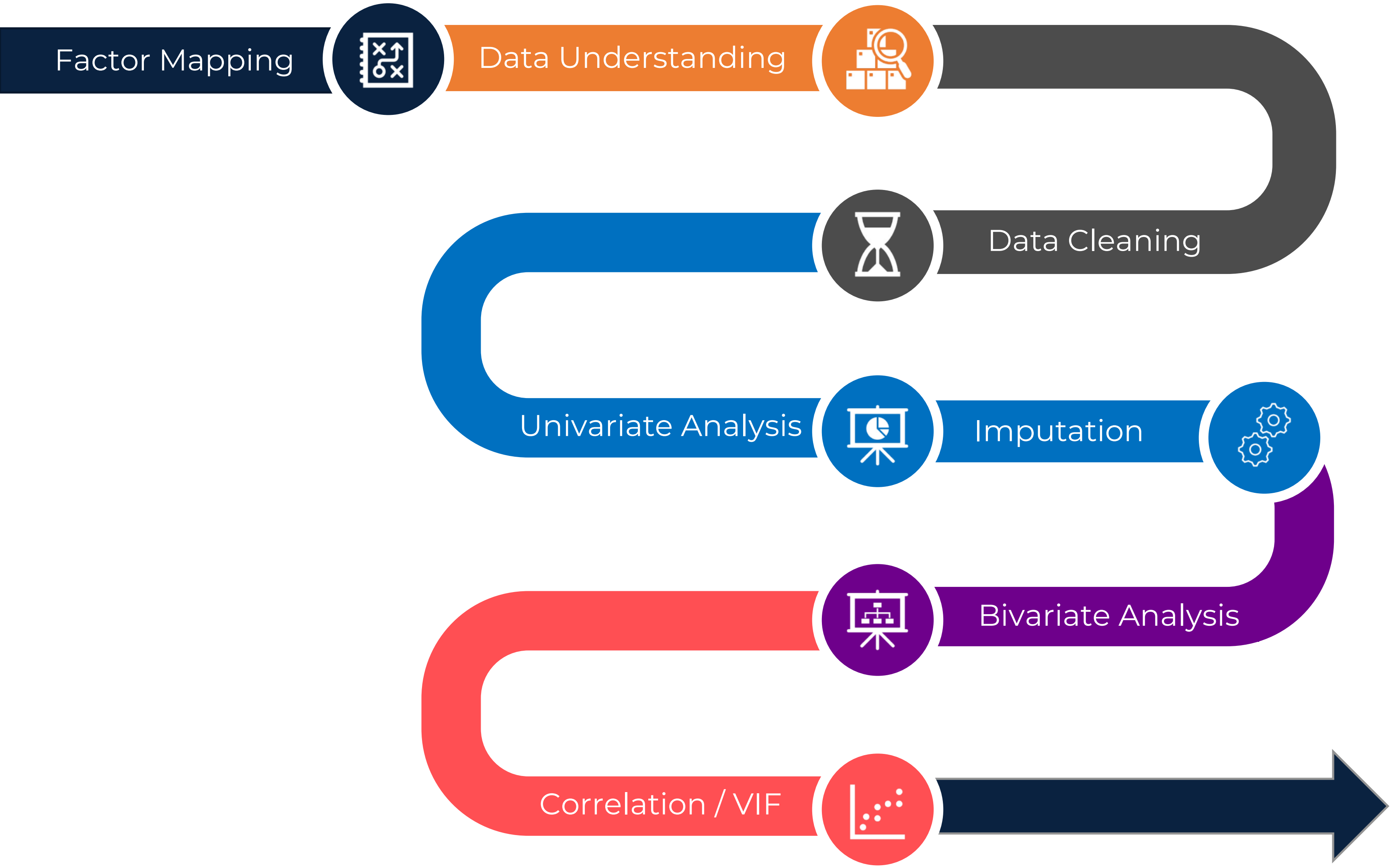
- **Characteristics** or factors with which a customer can be deemed profitable
- **Framework** to identify profitable customers to target for X credit cards

CONGLOMERATE



Customers of Interest
Customers who use X credit cards for payments at Y Cinemas

OUR PROCESS - RECAP



Factor Mapping
- Brainstormed possible factors
- Framed hypotheses

Data Understanding
- Created data dictionary
- Summarized dataset

Data Cleaning
- Preliminary preprocessing

Univariate Analysis
- Understanding data variables
- Outlier identification
- Imputed missing values

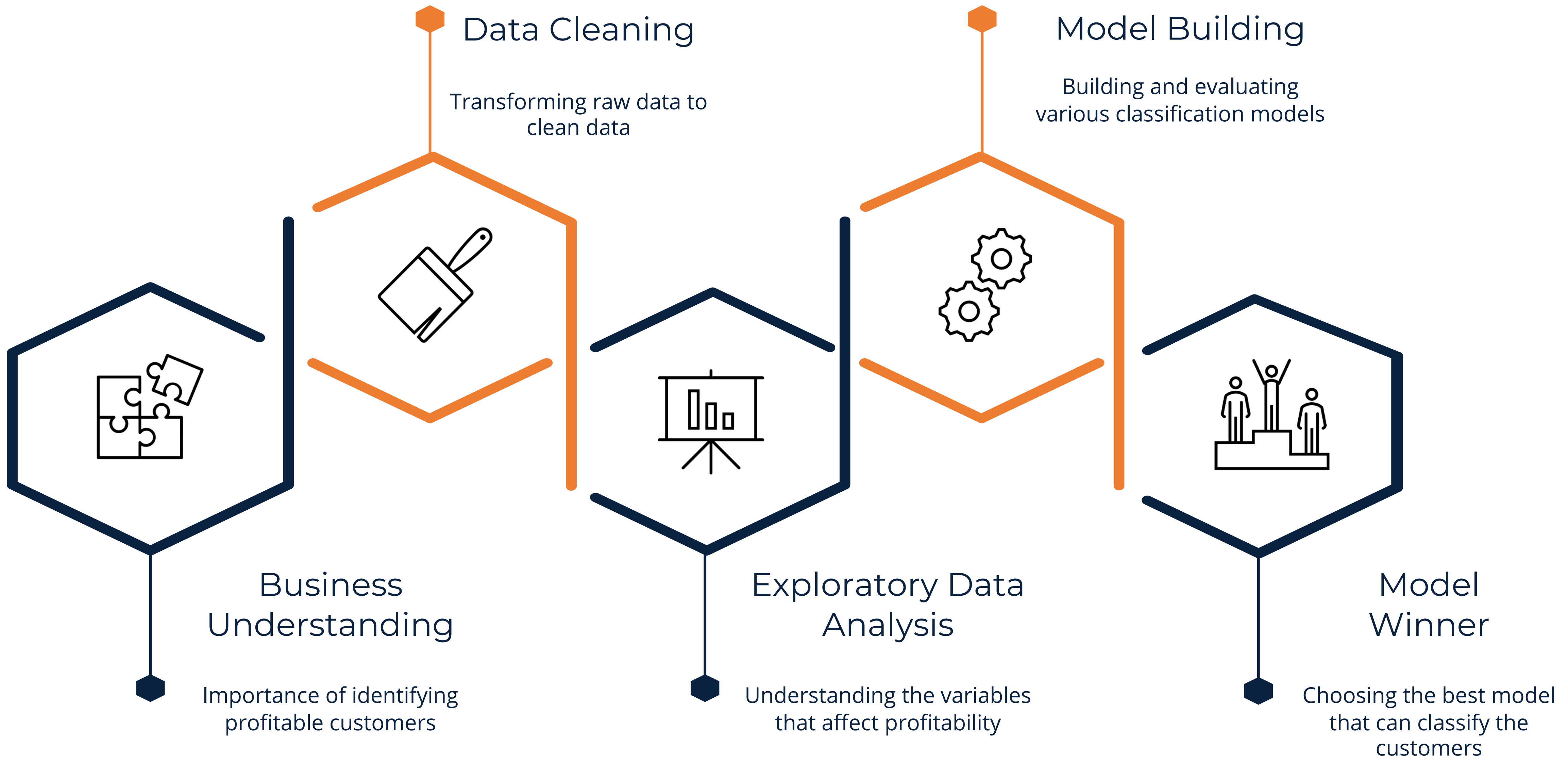
Bivariate Analysis
- Relationship b/w. data variables
- Validating hypotheses

Correlation/ VIF*
- Generated correlation matrix
- VIF iterations

Final Features to be fed in the model

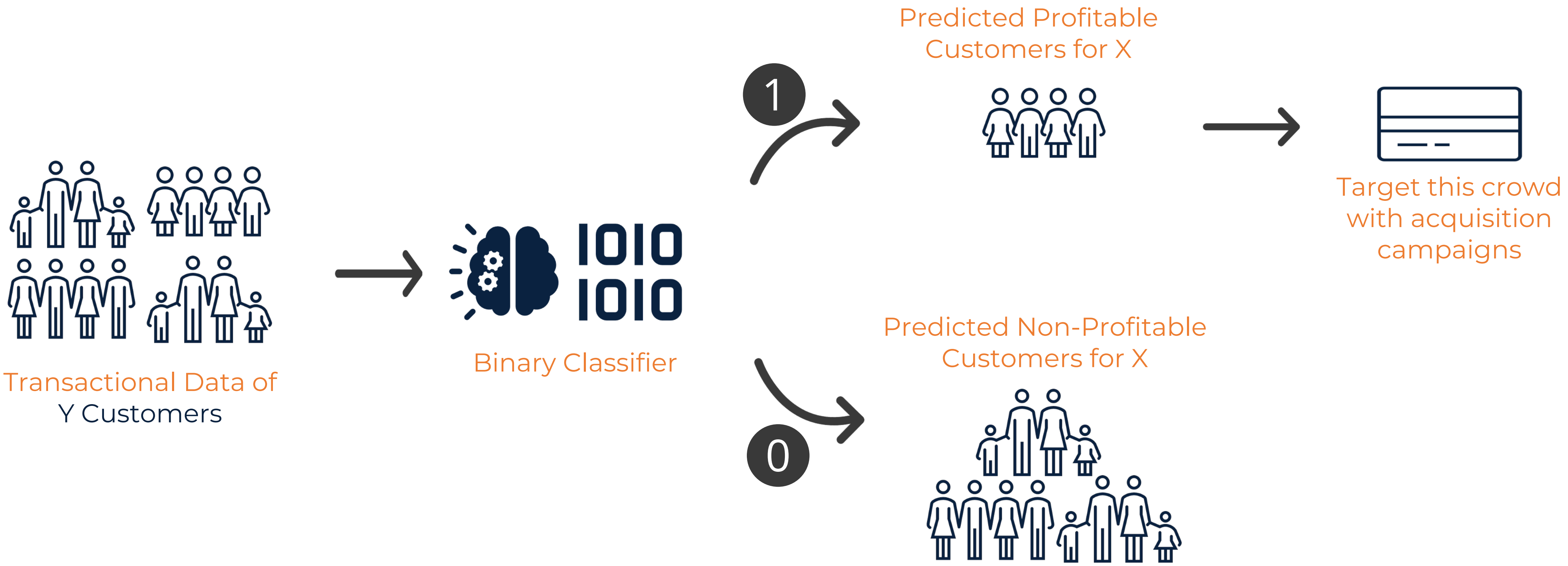
*Variable Inflation Factor

PROJECT LIFECYCLE



SOLUTION DESIGN

SOLUTION OVERVIEW FOR CROSS-SELLING X CREDIT CARDS TO Y CUSTOMER BASE



LEVERAGING THE EXISTING X CUSTOMER BASE AT Y FOR OUR ANALYSIS

Transactional Data of
X Customers
visiting Y cinemas

Existing customers are split based on
Profitability criteria provided by X Cards*

Revenue > = AED 350

~23 out of 100
Customers are Profitable

Revenue < AED 350

Analyze the
Characteristics of
Profitability

Final Feature Selection

*X Credit Cards called out their Profitable Customers to be generating an overall revenue of >= AED 350

CORRELATION & VIF ANALYSIS

USING VIF AND CORRELATION TO SELECT FEATURES

Removing data variables with a very high VIF >10 or correlation coefficient > 0.85 and iterating until we get satisfactory results

First Iteration

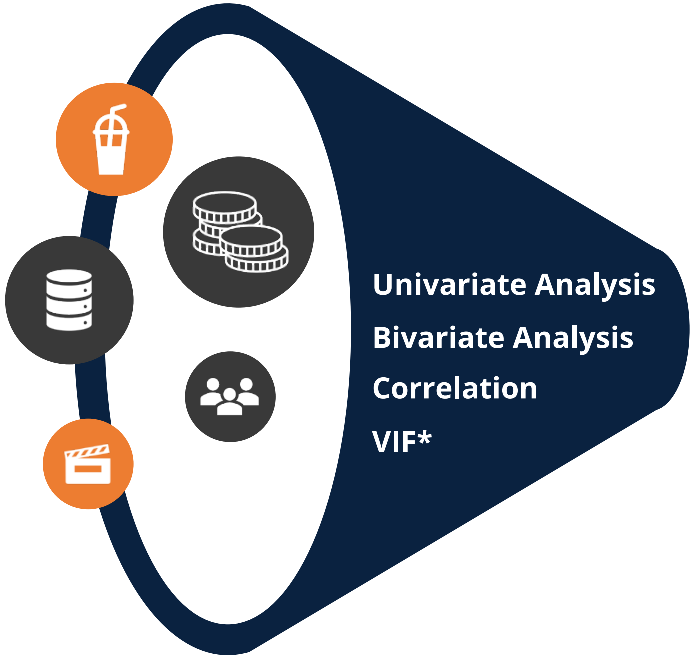
Sno	Features	VIF value
1	Last_60_days	22,687.12
2	Last_30_days	11,493.22
3	Last_90_days	11,222.60
4	Overall_Ticket_Amt	3,524.94
5	Booked_Amt	2,908.54
6	Overall_Spend	606.21
7	#Tickets	545.11
8	Pref_cinema_experience_#Ticket	454.55
9	Booked_Rdmption	111.43
10	Pref_movie_country_name_Spend	76.22
11	Pref_transaction_channel_Spend	46.64
12	Pref_transaction_channel_#Ticket	45.90
13	Pref_cinema_experience_Spend	43.99
14	#Movies_Watched	41.63
15	#Unique_Movies	41.21
16	Tickets_Weekend	40.80
17	Pref_movie_country_name_#Ticket	40.65

Sno	Features	VIF value
18	Pref_genre_name_Spend	27.75
19	#Weekends	24.75
20	Pref_film_rating_#Ticket	19.90
21	Pref_cinema_location_#Ticket	19.28
22	Pref_genre_name_#Ticket	18.08
23	Pref_cinema_location_Spend	14.281
24	Avg.Movie_Dur	8.87
25	Pref_film_rating_Spend	7.22
26	Avg_Tickt_Cost	5.79
27	Overall_FB_Spent	5.35
28	Is_internet_flag	3.63
29	Is_Action_flag	2.75
30	Is_mobile_flag	2.60
31	Is_Hollywood_flag	2.07
32	REVENUE_NAJM	1.70
33	New_Customer	1.52
34	Avg_Booking_Time	1.39

Citing Katrutsa and Strijov (2017) & Curto and Pinto (2011)

RELEVANT FEATURES WERE SELECTED AFTER FEATURE GENERATION

53
Variables



19
Features

Final List of Features

1. *# of Tickets bought on Weekends*
2. *Booking Amount*
3. *Booking Redemption*
4. *Average Movie Duration*
5. *Average Ticket Cost*
6. *Transaction Channel (Internet Ticketing)*
7. *Transaction Channel (Mobile Phone)*
8. *Watched an action movie or not*
9. *Watched a Hollywood movie or not*
10. *Amount spent on preferred cinema location*
11. *Amount spent on preferred film rating*
12. *Amount spent on preferred cinema experience*
13. *Amount spent on Food & Beverages*
14. *# of Unique Movies Watched*
15. *# of Visits in Last 90 days*
16. *Average time taken to make a booking*
17. *# of Tickets bought on Weekdays*
18. *Average Spend per Visit*
19. *Customer Tenure*

- 16 Features were selected from an exhaustive list of 53 variables through analysis
- 3 New Features were created from the existing features:
  - Customer Tenure, Average Spend per Visit, # of Tickets bought on Weekends

WORKFLOW EMPLOYED TO ACHIEVE THE DESIRED OUTCOME

Data Preparation

- Data cleaning
- Imputation
- EDA
- Outlier Treatment

Feature Engineering

- Correlation
- VIF
- Feature Generation
- Feature Scaling

WORKFLOW EMPLOYED TO ACHIEVE THE DESIRED OUTCOME

Data Preparation

- Data cleaning
- Imputation
- EDA
- Outlier Treatment

Feature Engineering

- Correlation
- VIF
- Feature Generation
- Feature Scaling

Build Classification Models

- Logistic Regression
- Naïve Bayes
- K-Nearest Neighbors
- Random Forest

Iterative Model Updating

Evaluate

Compare

Rebuild

Monitor

Deploy best performing model

RELEVANT MODELS FOR OUR PROBLEM STATEMENT - RECAP

	K –Nearest Neighbours	Logistic Regression	Support Vector Machine	Decision Tree	Boosting Techniques	Random Forest
Outliers	SENSITIVE	SENSITIVE	ROBUST	ROBUST	SENSITIVE	ROBUST
Collinearity	SENSITIVE	SENSITIVE	SENSITIVE	ROBUST	ROBUST	ROBUST
Performance	LOW	LOW	MEDIUM	MEDIUM	HIGH	HIGH

- As our dataset contains a high number of features, one decision tree cannot perform well and give accurate predictions
- The decision tree might overfit the training data, if the parameters are not well tuned
- This can be overcome if we use ensemble learning methods, like Random forest & Boosting because it will build N number of decision trees and give the outcome based on polling
- Random forest and boosting is a combination of many decision trees thus, more robust than a single decision tree
- Random Forest can be best suited for our dataset as it is not sensitive to outliers

ALGORITHMS IMPLEMENTED

Logistic Regression
Naïve Bayes
K-Nearest Neighbours
Random Forest

METRICS USED FOR MODEL EVALUATION

Area Under Curve - AUC
(Area under ROC* curve)
How good the classifier is at distinguishing between the profitable and non-profitable customers

Recall
(Capture Rate)
Fraction of customers which are correctly identified as profitable out of all actual profitable customers

Precision
(Conversion Rate)
Fraction of customers who turn out to be profitable among all the *predicted* profitable customers

- Recall is low ⇒ Model will classify profitable customers as non-profitable
- Precision is low ⇒ Model will classify non-profitable customers as profitable

“ The cost of lower recall is way higher than the cost of lower precision ”

(As per business requirements)

Why Accuracy is *not* our evaluation metric?

Not a good measure of classifier performance for highly imbalanced dataset

(In our case, distribution of majority-to-minority class is 77:23, then labelling all data points as majority class would give you 77% accuracy which is really good score, but in fact the model has not learned anything)

*Receiver Operating Characteristics

Citing Wasikowski and Chen (2010)

MODEL PERFORMANCE RESULTS

Model	AUC	Recall	Precision	Threshold
Logistic Regression	55%	69%	27%	23%
Naïve Bayes	54%	65%	27%	4%
K-Nearest Neighbors	54%	80%	26%	19%
Random Forest	58%	78%	28%	43%

LOGISTIC REGRESSION AND NAÏVE BAYES ARE NOT SUITED FOR OUR PROBLEM

Model	AUC	Recall	Precision	Threshold
Logistic Regression	55%	69%	27%	23%
Naïve Bayes	54%	65%	27%	4%
K-Nearest Neighbors	54%	80%	26%	19%
Random Forest	58%	78%	28%	43%

1 Sensitive to Outliers

- Outlier Treatment is required for good model performance
- Treating the outlier leads to losing the actual data

2 Lower Recall

- Model would miss out on the actual profitable customers

KNN IS THE CHALLENGER MODEL

Model	AUC	Recall	Precision	Threshold
Logistic Regression	55%	69%	27%	23%
Naïve Bayes	54%	65%	27%	4%
K-Nearest Neighbors	54%	80%	26%	19%
Random Forest	58%	78%	28%	43%

KNN vs Random Forest

- KNN has a recall of 80% while Random Forest has 78%
- Comparing the AUC scores, Random Forest is higher with a score of 58% while KNN has 54%
- Results might be inflated for KNN model since we used KNN Imputer for imputing

KNN IS THE CHALLENGER MODEL

Model	AUC	Recall	Precision	Threshold
Logistic Regression	55%	69%	27%	23%
Naïve Bayes	54%	65%	27%	4%
K-Nearest Neighbors	54%	80%	26%	19%
Random Forest	58%	78%	28%	43%

KNN vs Random Forest

- KNN has a recall of 80% while Random Forest has 78%
- Comparing the AUC scores, Random Forest is higher with a score of 58% while KNN has 54%
- Results might be inflated for KNN model since we used KNN Imputer for imputing

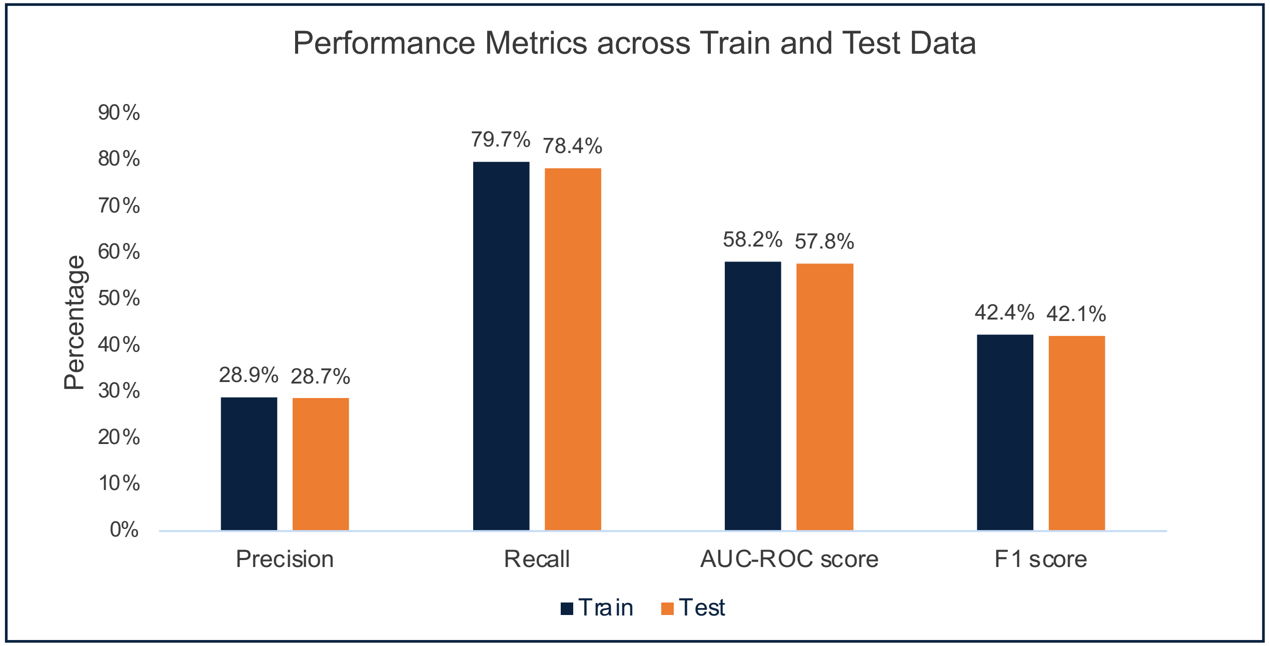
We trade-off 2% decrease of the recall for a 4% increase in AUC score

RANDOM FOREST WAS SELECTED AS THE CHAMPION MODEL BASED ON MULTIPLE ITERATIONS

Confusion Matrix for Test Data

Predicted \ Actual	Profitable	Not Profitable
Profitable	861	2,130
Not Profitable	237	1,270

Probability Threshold: 43%



AUC Score

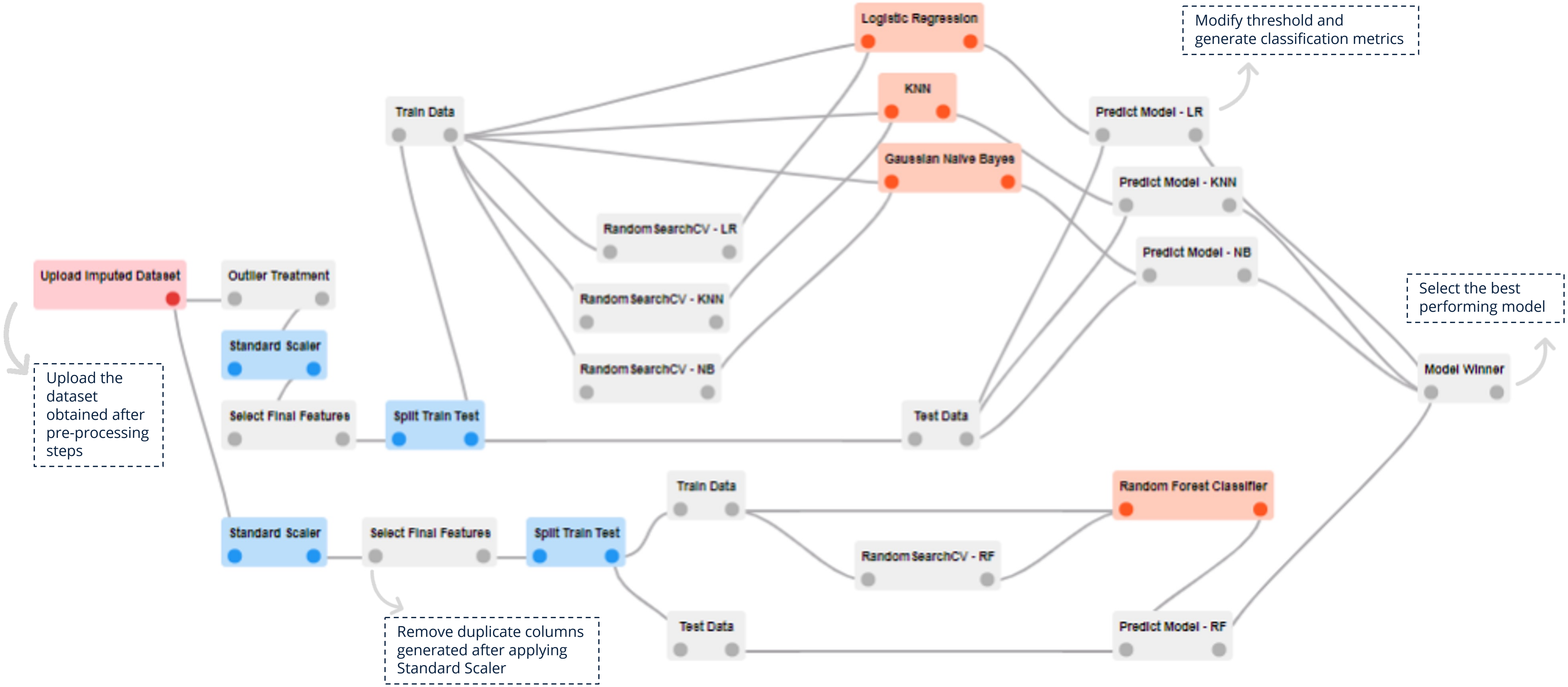
Recall

1. **Not Overfitted**
   The difference between training and testing metrics is negligible

2. **High Recall**
   signifies that it is better at predicting actual profitable customers

3. **High AUC Score**
   signifies the model is better at distinguishing between profitable and non-profitable customers

CO.DX BLUEPRINT FOR MODELLING



BUSINESS IMPACT

BUSINESS IMPACT

Understand and leverage the existing customer base

Framework for identifying profitable customers to target

Cross selling NAJM credit cards to other business units

Conventional Marketing

↓

Data-Driven Marketing

The model manages to capture 78% of the profitable customers which is our target segment

Reduction in marketing cost and increase in ROI*

Tailoring acquisition campaigns for profitable customers

*ROI – Return on Investment

NEXT STEPS

Enhance Model Performance

- Fine tuning the Model results(increase capture and conversion rate)
- Getting new data features from the client – Frequency of visiting, Cancellation, Offers provided
- Generate other possible features from the existing features

Propensity Model

- Build a framework to score customers' propensity
- Marketing materials and campaigns can be tailored to individuals based on their estimated propensity to purchase

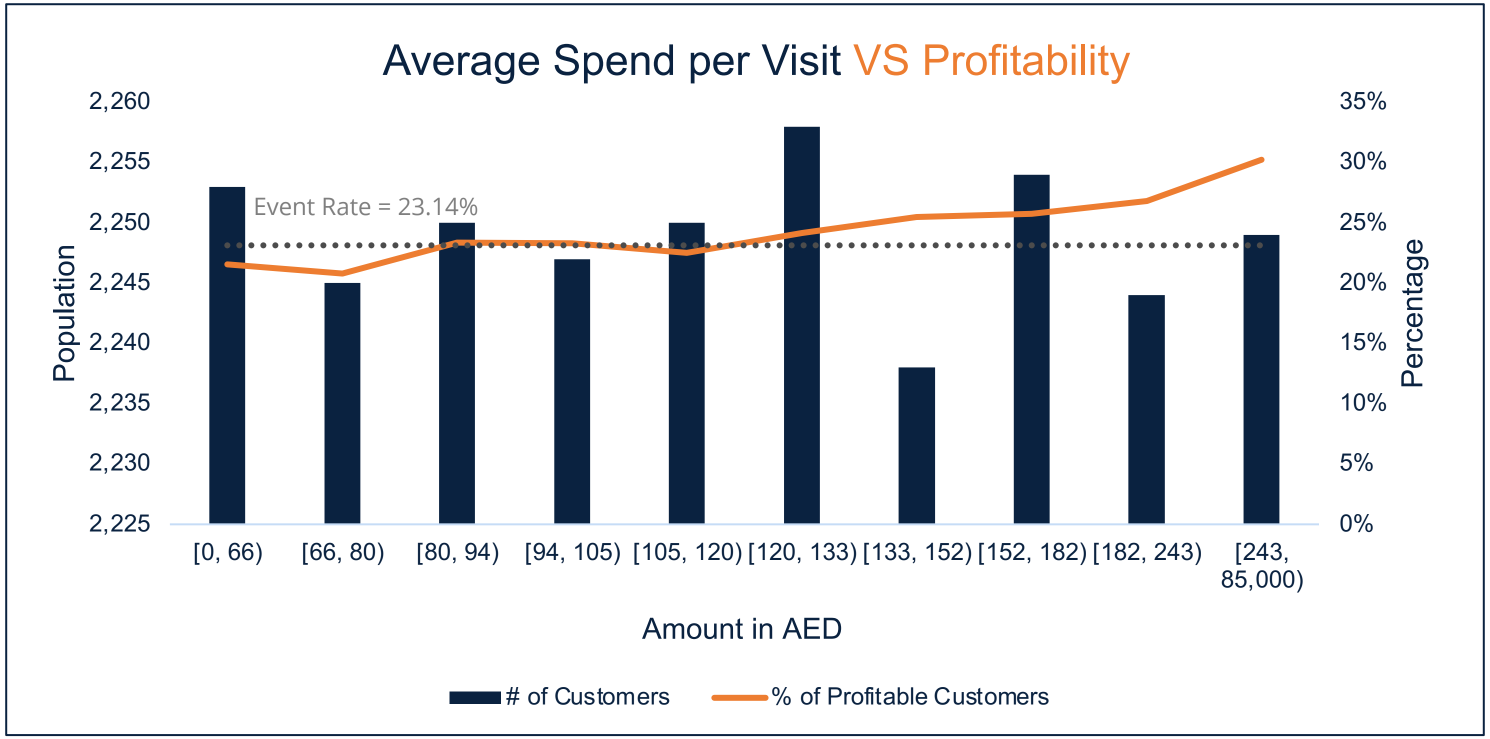
Tool for Marketing Team

- Tool which recommends acquisition strategies tailored specifically for a customer

THANK YOU

APPENDIX

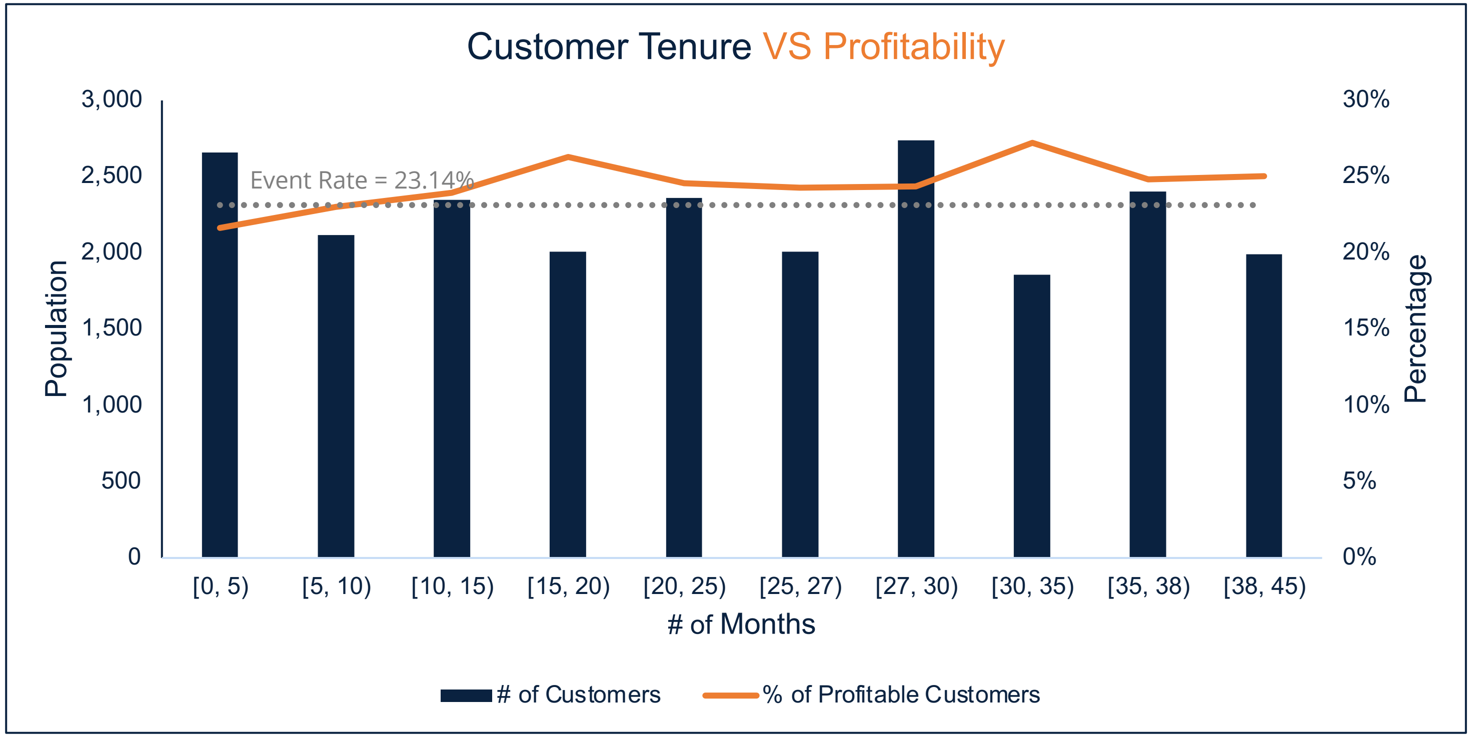
CUSTOMERS SPENDING MORE THAN AED 120 PER VISIT ARE PROFITABLE



Observations:

- As average spend per visit increases, profitability increases
- Customers spending more than AED 243 are highly profitable

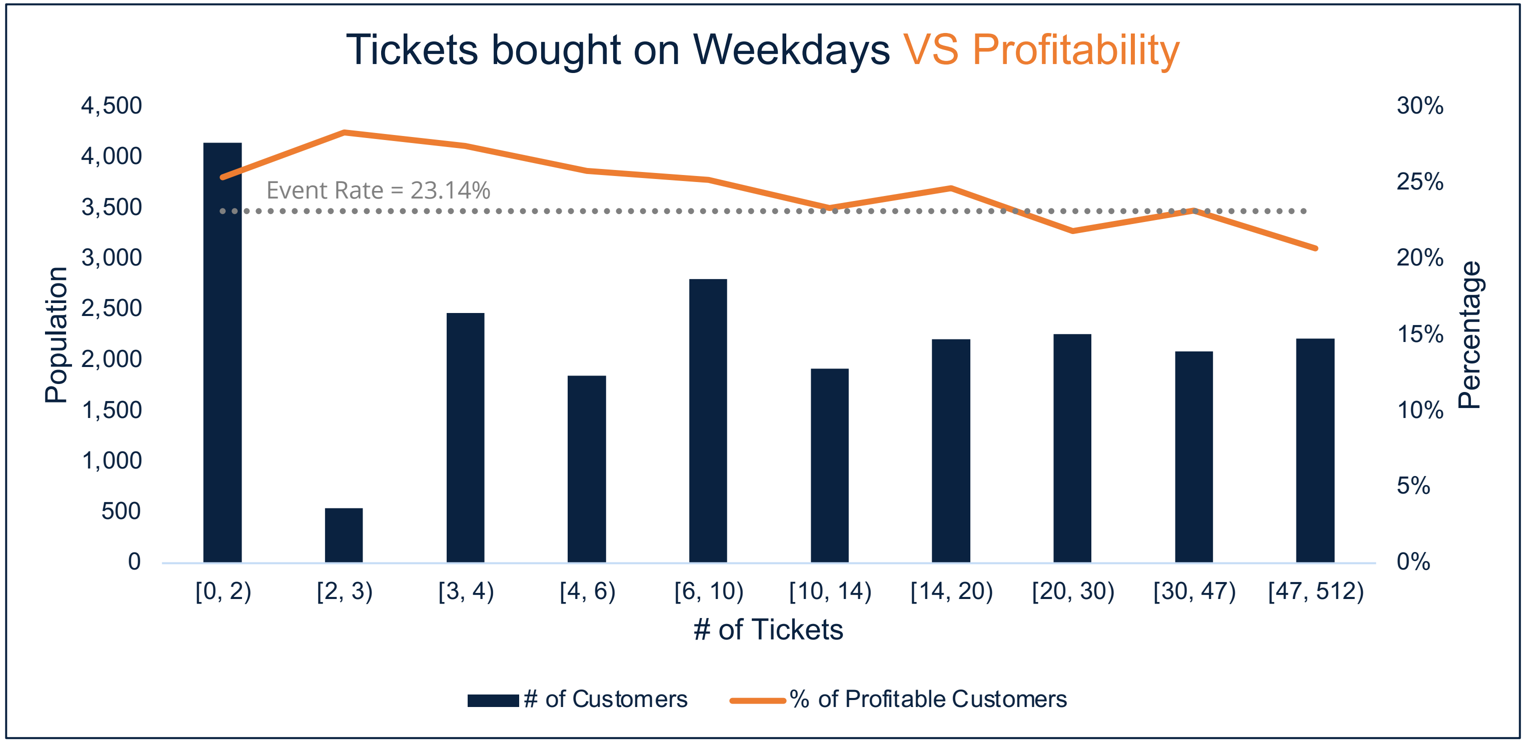
CUSTOMERS WITH A TENURE OF 10 MONTHS AND MORE ARE PROFITABLE



Observations:

- Spikes in profitability is observed when the customer tenure is between 15-19 & 30-34 months
- Profitability flattens as a customer approaches two years of tenure but increases again as it approaches the third year

PROFITABILITY DECREASES, AS NUMBER OF TICKETS BOUGHT ON WEEKDAY INCREASES



Observations:

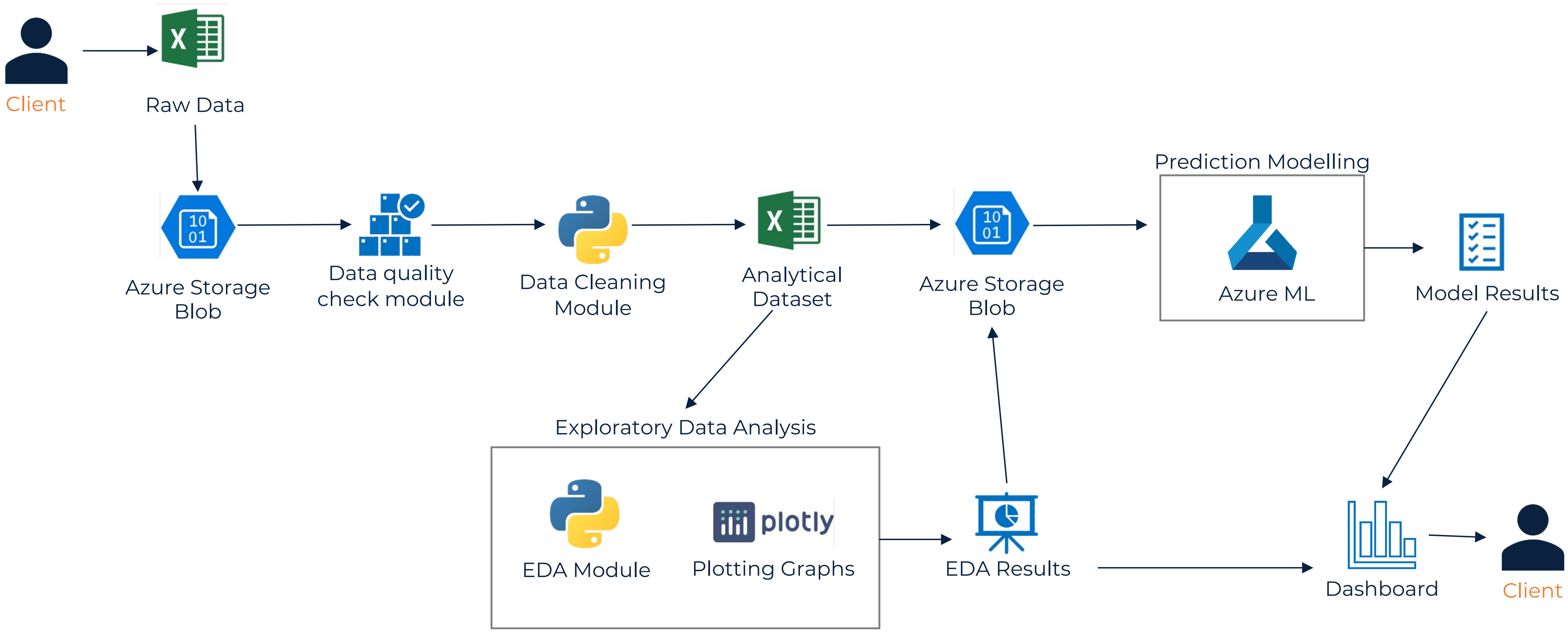
- ~18% people buy up to 1 ticket on weekdays
- People buying only 2 tickets during weekdays are highly profitable

RELEVANT MODELS FOR OUR PROBLEM STATEMENT

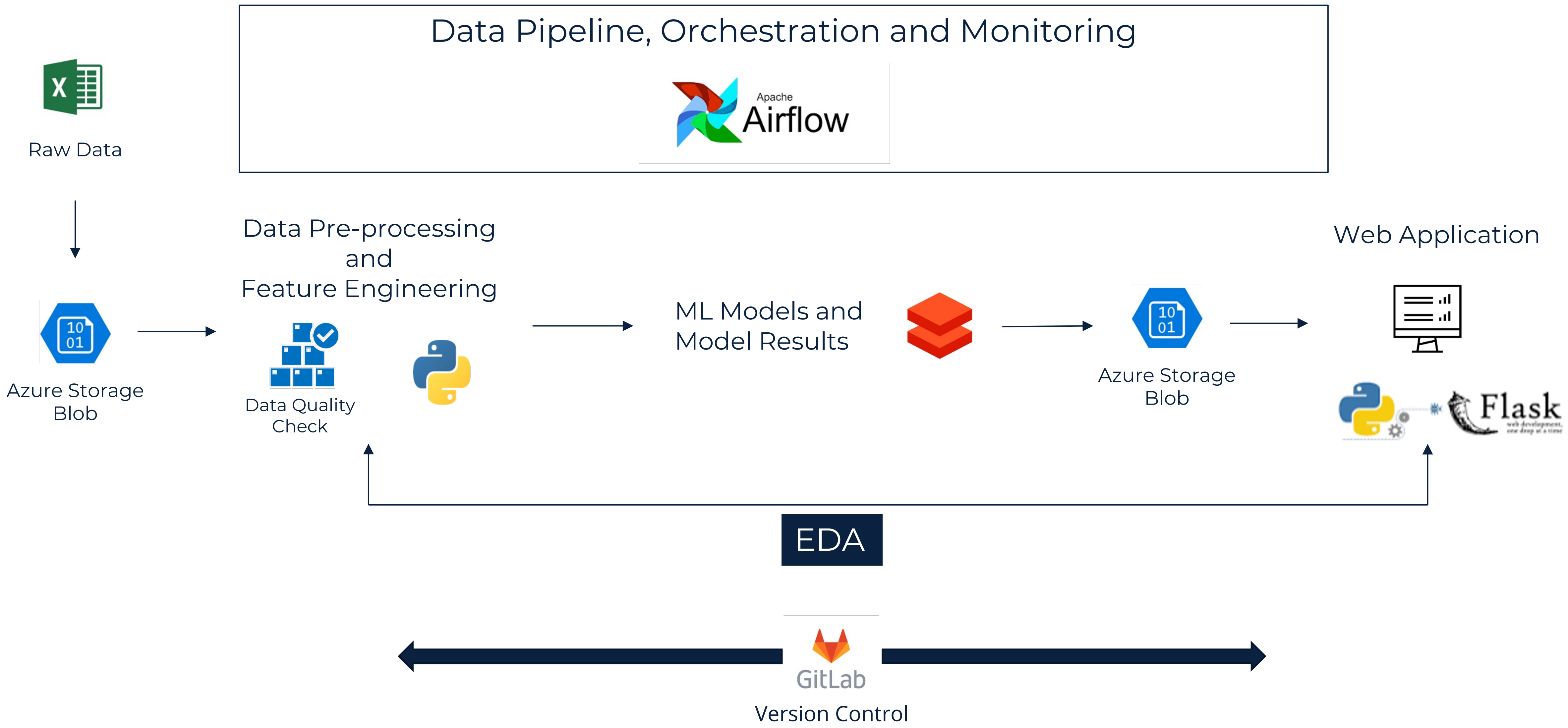
	K –Nearest Neighbours	Logistic Regression	Support Vector Machine	Decision Tree	Boosting Techniques	Random Forest
Outliers	SENSITIVE	SENSITIVE	ROBUST	ROBUST	SENSITIVE	ROBUST
Collinearity	SENSITIVE	SENSITIVE	SENSITIVE	ROBUST	ROBUST	ROBUST
Performance	LOW	LOW	MEDIUM	MEDIUM	HIGH	HIGH

- As our dataset contains a high number of features, one decision tree cannot perform well and give accurate predictions
- The decision tree might overfit the training data, if the parameters are not well tuned
- This can be overcome if we use ensemble learning methods, like Random forest & Boosting because it will build N number of decision trees and give the outcome based on polling
- Random forest and boosting is a combination of many decision trees thus, more robust than a single decision tree
- Random Forest can be best suited for our dataset as it is not sensitive to outliers

ARCHITECTURE DIAGRAM



ARCHITECTURE DIAGRAM USING AIRFLOW



FEATURE IMPORTANCE

Feature	Importance
Avg.Movie_Dur	0.100
Booked_Rdmption	0.098
Spend_per_movie	0.093
Avg_Booking_Time	0.090
Avg_Tickt_Cost	0.087
Customer_Tenure	0.067
Booked_Amt	0.066
Pref_film_rating_Spend	0.062
Pref_cinema_location_Spend	0.062
Pref_cinema_experience_Spend	0.061
#Weekday_Tickets	0.050
#Unique_Movies	0.034
Last_90_days	0.031
#Weekends	0.028
Overall_FB_Spent	0.023
Is_Hollywood_flag	0.019
Is_Action_flag	0.011
Is_internet_flag	0.011
Is_mobile_flag	0.007