ELITE - **E**nhancing **L**oyal **I**n-house **T**arget **E**xpansion

A case study for Maybank

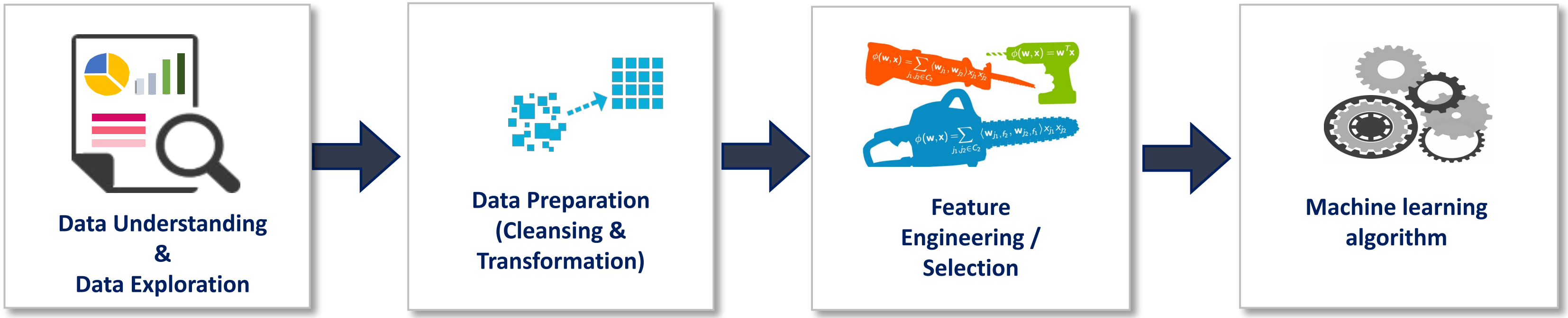
Agenda

- ✓ Introduction
- ✓ Data Analysis
- ✓ Modelling Approach
- ✓ Model Results and Interpretation
- ✓ Business Recommendations
- ✓ Model Monitoring
- ✓ MLOps Overview
- ✓ Conclusion

Introduction

- ✓ The case study aims to identify the potential affluent customers within the bank's existing customer base.
- ✓ Upgrading existing-to-bank (ETB) customers to affluent status is expected to boost revenue and foster stronger customer relationships.
- ✓ Through data analysis, the project seeks to uncover affluent behavior patterns, enabling targeted marketing strategies.
- ✓ Employing Logistic Regression, Random Forest Classifier, Extra Trees Classifier, and MLP Model, the project compares their efficacy in achieving the segmentation upgrade.

Approach



$\phi(\mathbf{w},\mathbf{x}) = \sum_{j_1,j_2 \in C_2} \langle \mathbf{w}_{j_1}, \mathbf{w}_{j_2} \rangle x_{j_1} x_{j_2}$
$\phi(\mathbf{w},\mathbf{x}) = \mathbf{w}^T\mathbf{x}$
$\phi(\mathbf{w},\mathbf{x}) = \sum_{j_1,j_2 \in C_2} \langle \mathbf{w}_{j_1,f_2}, \mathbf{w}_{j_2,f_1} \rangle x_{j_1} x_{j_2}$

Data Understanding
&
Data Exploration

Data Preparation
(Cleansing &
Transformation)

Feature
Engineering /
Selection

Machine learning
algorithm

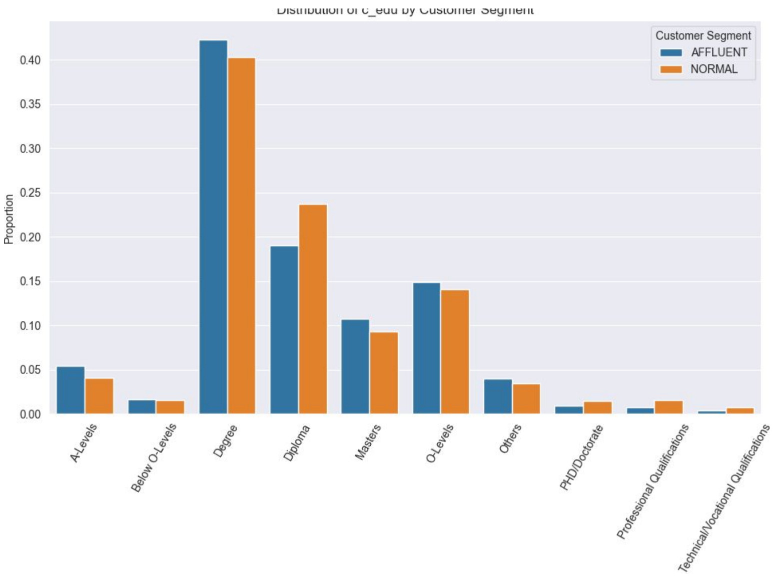
Data Understanding & Exploration

Data Understanding
&
Data Exploration

Data Analysis

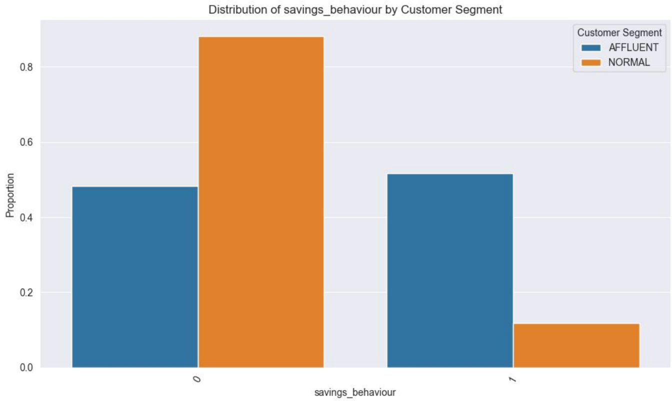
Data Description:

- Total number of records: 66083
- Out of 66K records 16.5 % customers are affluent
- Based on the C_ID, 47857 unique customers are available
- % of customers with one record in the dataset 69.94 %
- % of customers with multiple records for the same customer ID's 30.06 %
- Removed duplicate rows for customers who have multiple records with the same customer ID, keeping only one instance of the repeated rows.



- ~ 75% of customer base has a undergraduation

Data Analysis

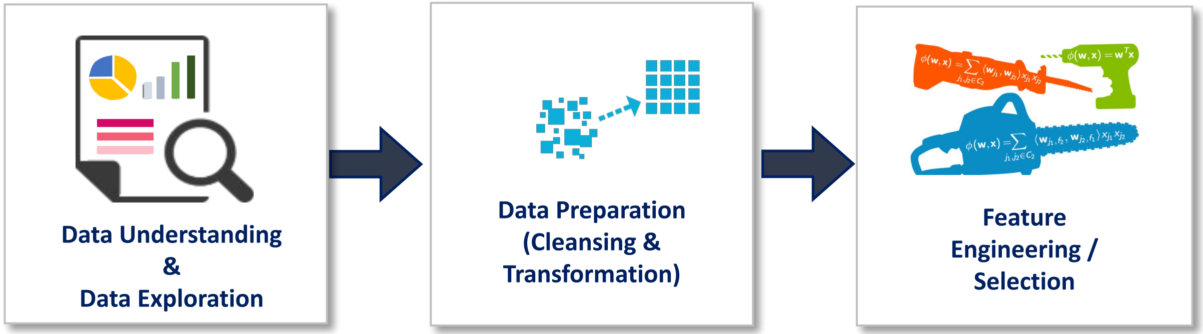


~ 40% of Affluent customers has a positive balance in both CASA and TD accounts

Top 25% of Affluent customers asset value is 17x times greater than the normal customers

C_seg	Mean Asset Value	Median Asset Value	75th Percentile
AFFLUENT	120523.4963	76192.82	145003.42
NORMAL	19336.58944	207.77	8486.75

Data Preparation


$\phi(\mathbf{w},\mathbf{x}) = \sum_{j_1,j_2 \in C_2} \langle \mathbf{w}_{j_1}, \mathbf{w}_{j_2} \rangle x_{j_1} x_{j_2}$
$\phi(\mathbf{w},\mathbf{x}) = \mathbf{w}^T\mathbf{x}$
$\phi(\mathbf{w},\mathbf{x}) = \sum_{j_1,j_2 \in C_2} \langle \mathbf{w}_{j_1,f_2}, \mathbf{w}_{j_2,f_1} \rangle x_{j_1} x_{j_2}$

Data Understanding
&
Data Exploration

Data Preparation
(Cleansing &
Transformation)

Feature
Engineering /
Selection

Feature Engineering

Added several new features to capture different aspects of the customer behavior, financial status and engagement with bank products.

S.No	Feature Name	Feature Description
1	age_bin	This feature converts the customer age into categorical age groups. If C_AGE <18 : 'Minor' If 18 <= C_AGE < 30 : 'Young Adult' If 30 <= C_AGE < 60 : 'Adult' If C_AGE >= 60 : 'Senior'
2	wealth_accumulation	Combine features related to asset value, account balances, and investment values to calculate a metric representing the customer's overall wealth accumulation rate. Wealth Accumulation = Asset_value / C_AGE
3	monthly_txn_frequency	Instead of using absolute transaction amounts, calculate the number of transactions per month to capture spending habits more accurately. Transaction Frequency = ANN_N_TRX / 12
4	credit_utilization	Calculate the ratio of average credit card balance to credit card limit to capture the customer's credit utilization behavior. Credit Utilization Ratio = CC_AVE / CC_LMT

Feature Engineering

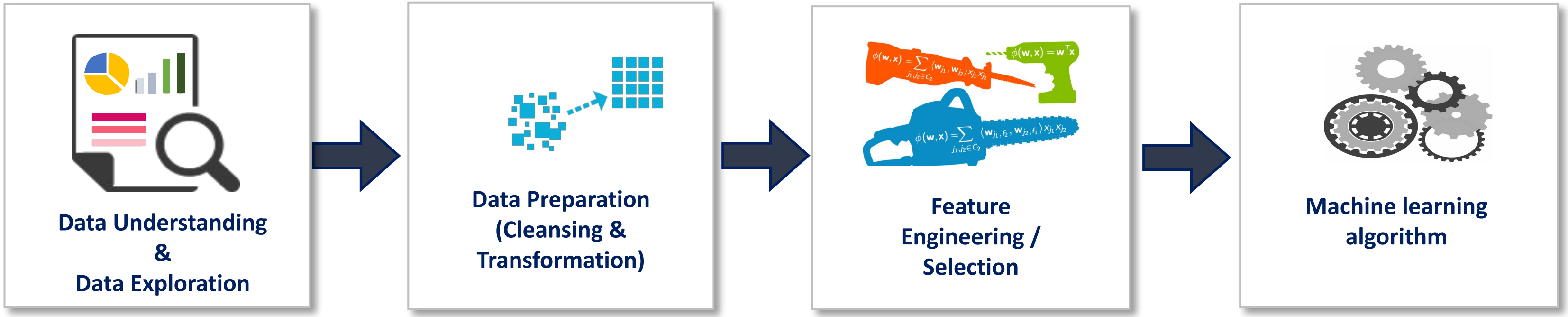
Added several new features to capture different aspects of the customer behavior, financial status and engagement with bank products.

S.No	Feature Name	Feature Description
5	savings_behaviour	Savings Behavior Indicator: Binary indicator based on whether the customer has a positive balance in both CASA and TD accounts. If MTHCASA > 0 and MTHTD > 0: 1 (indicating positive balance in both accounts) Else: 0 (indicating no positive balance in both accounts)
6	debt_to_asset_ratio	Debt-to-Asset Ratio: Ratio of the total debt (annual credit card transaction amount) to the total asset value. Formula: Debt_to_Asset_Ratio = ANN_TRN_AMT / Asset_value
7	txn_freq_per_prd	This feature calculates the transaction frequency per product by dividing the annual number of transactions by the number of distinct products held. Formula: Transaction_Frequency_per_Product = ANN_N_TRX / NUM_PRD
8	investment_to_debt_ratio	This feature calculates the ratio of the total investment value (average unit trust value) to the total debt (annual credit card transaction amount). Formula: Investment_to_Debt_Ratio = UT_AVE / ANN_TRN_AMT

Feature Selection

- Total of 29 features were provided
- After analysis removed the C_ID - Dummy Customer ID , PC - Dummy postal code
- Added additional 8 new intuitive feature
- Removed highly correlated and less informative features using recursive feature selection and elimination strategy
- Total of 31 features were selected out of 37 features
- Since the data provided is customer transaction and product data, it's not right to impute the missing values with standard statistical method
    - For numerical variables the missing values are imputed with 0
    - For categorical variables the missing values are imputed with ‘UNKNOWN’ keyword

Approach



$\phi(\mathbf{w},\mathbf{x}) = \sum_{j_1,j_2 \in C_2} \langle \mathbf{w}_{j_1}, \mathbf{w}_{j_2} \rangle x_{j_1} x_{j_2}$
$\phi(\mathbf{w},\mathbf{x}) = \mathbf{w}^T\mathbf{x}$
$\phi(\mathbf{w},\mathbf{x}) = \sum_{j_1,j_2 \in C_2} \langle \mathbf{w}_{j_1,f_2}, \mathbf{w}_{j_2,f_1} \rangle x_{j_1} x_{j_2}$

Data Understanding
&
Data Exploration

Data Preparation
(Cleansing &
Transformation)

Feature
Engineering /
Selection

Machine learning
algorithm

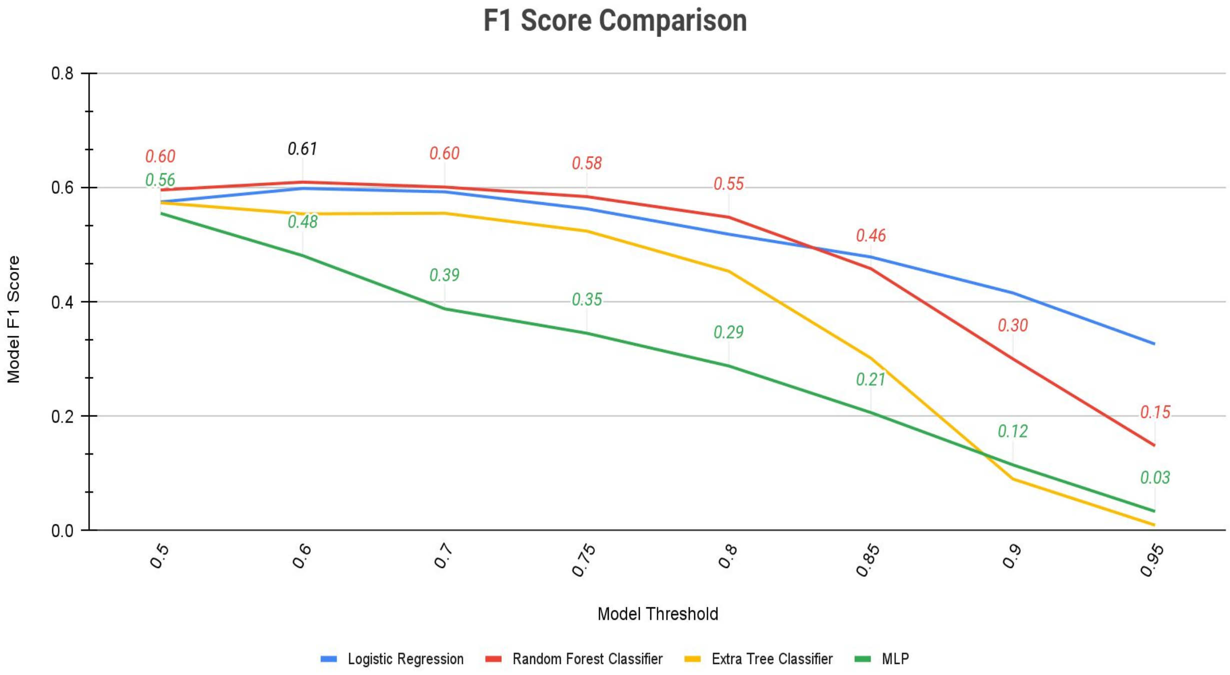
1. Logistic Regression
2. Extra Tree Classifier
3. MLP
4. Random Forest Classifier

Modelling Results

- Developed four models for this case study
- Out of 4 models it is evident that the Random Forest Classifier has outperformed others in terms of Recall, F1 Score, ROC AUC and Class Accuracy
- 28.17% uplift over the baseline model

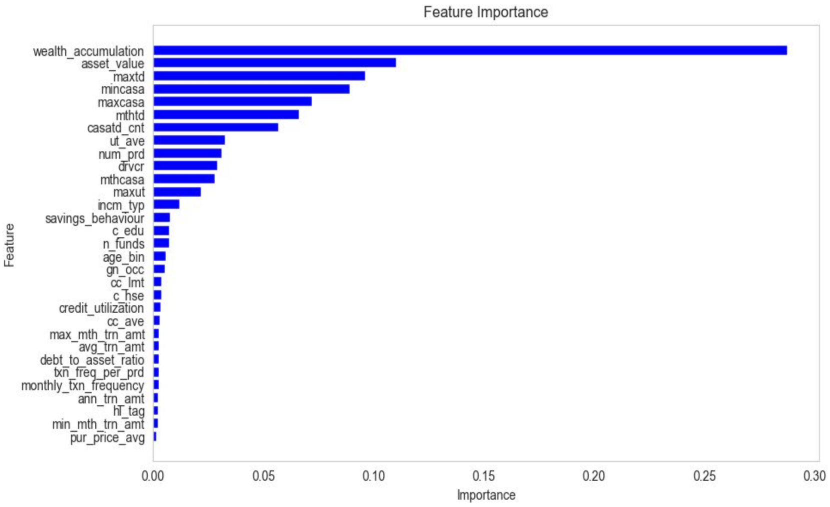
Model Name	Overall Accuracy	Accuracy: Normal	Accuracy: Affluent	Affluent Class Precision	Affluent Class Recall	Affluent Class F1 Score	ROC AUC
MLPClassifier	87.62%	95.71%	46.77%	**68.36%**	46.77%	0.555	0.87
ExtraTreesClassifier	82.16%	84.04%	72.68%	47.42%	72.68%	0.574	0.86
LogisticRegression	81.41%	82.45%	76.16%	46.22%	76.16%	0.575	0.86
RandomForestClassifier	82.50%	83.36%	**78.17%**	48.19%	**78.17%**	**0.596**	**0.88**

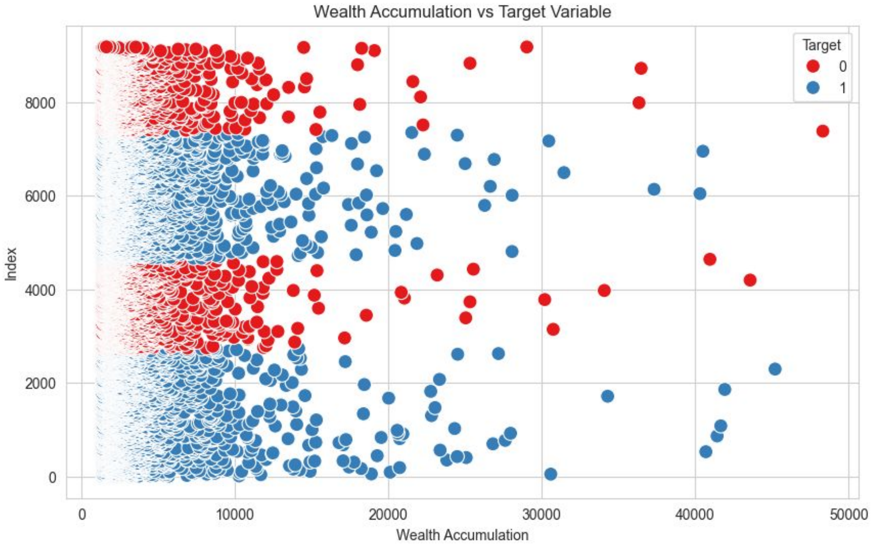
Modelling Results



- From the plot, it is suggested that choosing model threshold of **0.6** will result in highest F1 Score
- In the notebook, plotted precision, recall curve for different threshold
- Based on the business priority thresholds can be decided

Modelling Results



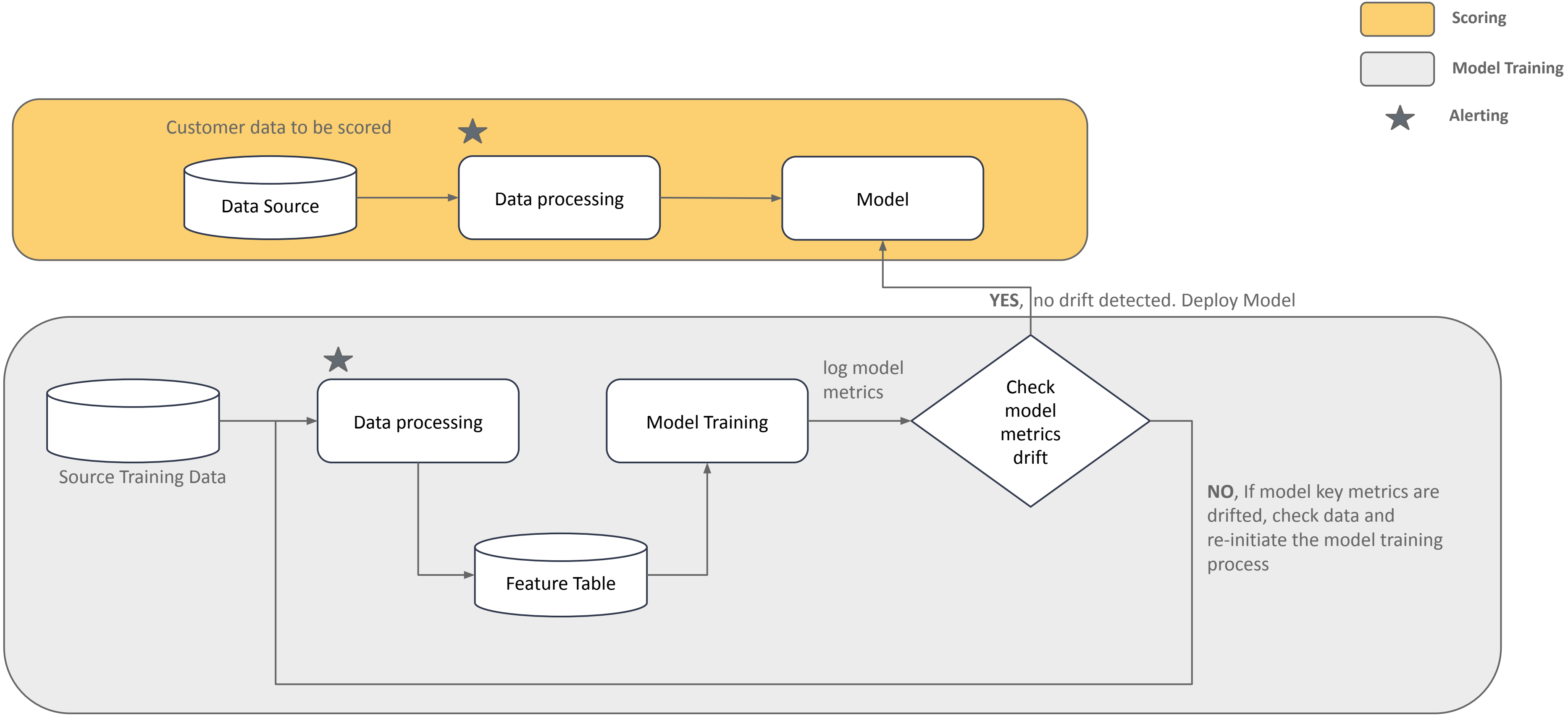


C_seg	mean	25%	50%	75%
AFFLUENT	2074.19	548.90	1363.64	2569.62
NORMAL	328.68	0.00	4.38	167.22

Customers with higher Asset value should be targeted for upgrading their segment.

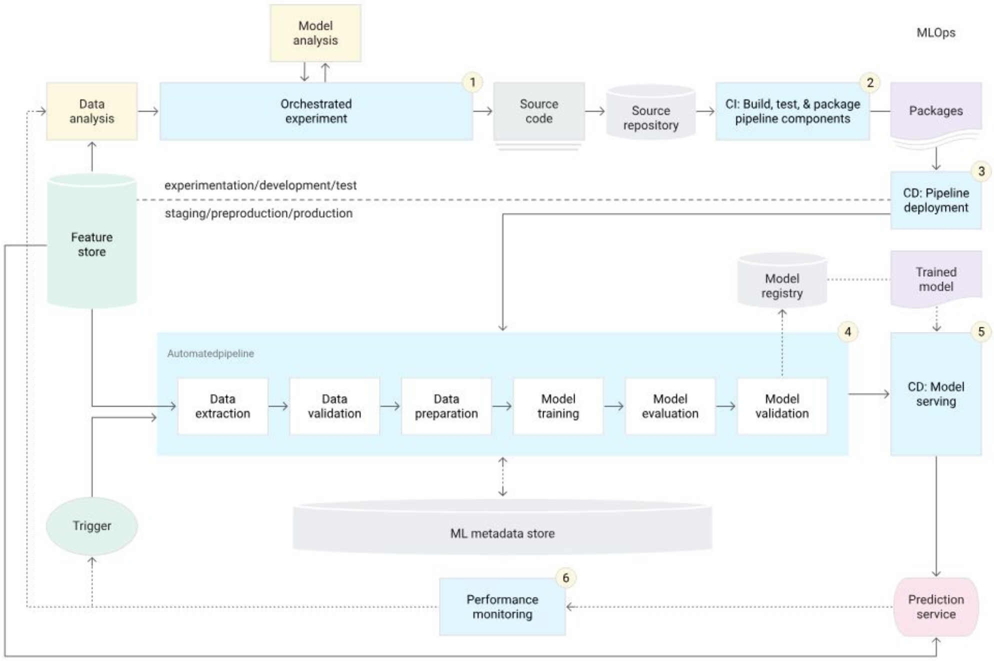
The 'AFFLUENT' segment exhibits significantly higher mean and median wealth accumulation (2074.19 and 1363.64, respectively) compared to the 'Normal' segment (328.68 and 4.38, respectively).

High level Model Flow & Monitoring



Scoring
Model Training
Alerting
Customer data to be scored
Data Source
Data processing
Model
YES, no drift detected. Deploy Model
Data processing
Model Training
log model metrics
Check model metrics drift
Source Training Data
Feature Table
NO, If model key metrics are drifted, check data and re-initiate the model training process

CI/CD and automated ML pipeline



Credits: CI/CD pipeline automation

1. **Development and experimentation** - Iteratively experiment with new ML algorithms and modeling techniques in a coordinated manner.

2. **Pipeline continuous integration**: Compile source code and conduct tests for deployment in subsequent stages.

3. **Pipeline continuous delivery**: Deploy artifacts generated by the CI stage to the target environment.

4. **Automated triggering**: Automatically initiate production processes based on schedules or triggers.

5. **Model continuous delivery**: Deploy trained models as prediction services for making predictions.

6. **Monitoring**: Gather real-time statistics on model performance using live data.

Conclusion & Improvements

Conclusion

- 4 Model were developed, out of that RF has outperformed for varying thresholds
- Wealth Accumulation is key feature to identify the next possible Affluent customers for upselling

Next Steps for Improvements:

- Data: There were columns with missing rate more 60%, Imputing missing values with 0 for numerical variables and 'UNKNOWN' for categorical variables is a common approach, especially when dealing with transaction and product data. Need to identify the root cause for the missing values. Also could consider imputation methods like K-nearest neighbors (KNN) imputation or using predictive models to estimate missing values.
- Downsampling of majority class
- Add more transactional features like: monthly transaction volume, transaction frequency, RFM features and details of the products they hold with the bank