E-Commerce & Retail B2B Case Study

BY: Ketan Chavda, Parth Sharma, Srijan Bajpai

PROBLEM STATEMENT

Schuster, a multinational sports retail company, grapples with vendor late payments, affecting its financial stability and vendor relationships. They issue invoices detailing transaction specifics, including payment due dates based on contractual terms. To tackle this issue, Schuster is turning to data science, aiming to comprehend customer payment patterns and predict late payments for open invoices. The objective is to streamline collections, minimize financial losses, and strengthen vendor relations by optimizing the payment process through data-driven insights and predictive analytics.

OBJECTIVE OF THE STUDY

- Schuster would like to better understand the customers' payment behaviour based on their past payment patterns (customer segmentation).
- Using historical information, it wants to be able to predict the likelihood of delayed payment against open invoices from its customers.
- It wants to use this information so that collectors can prioritise their work in following up with customers beforehand to get the payments on time.

ANALYSIS APPROACH

Data Understanding and Preprocessing:

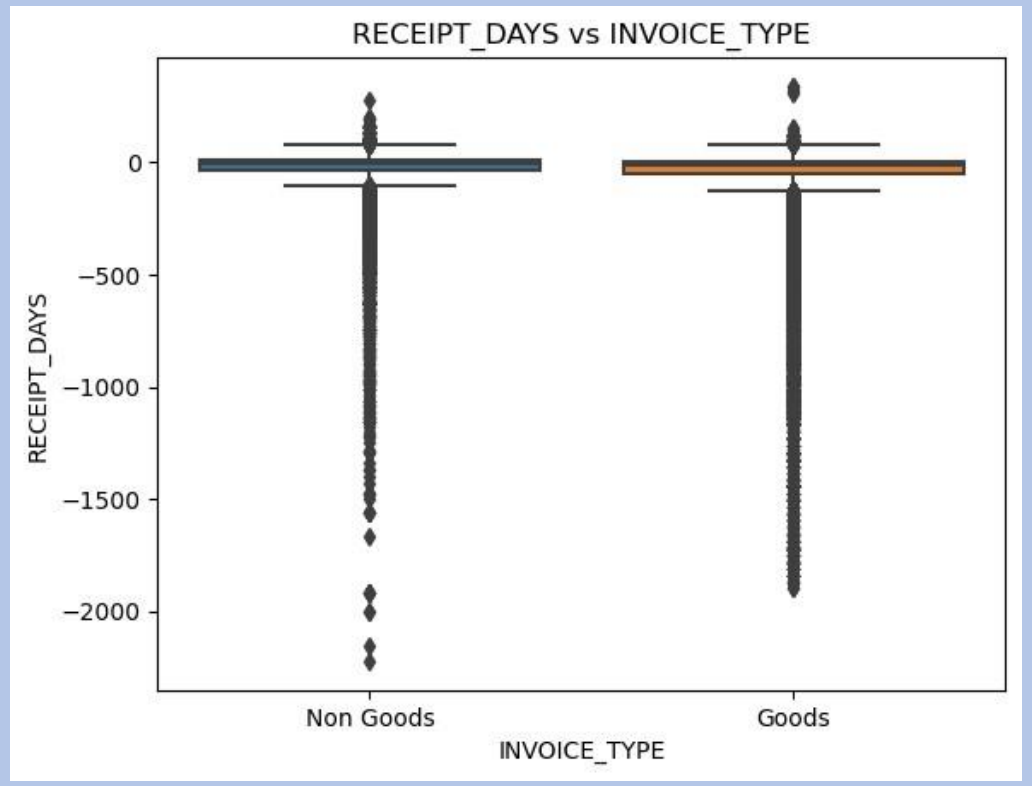
- Identified Schuster's business problem related to vendor payment delays and the need for datadriven solutions.
- Two sets of data were provided
- **RECEIVED PAYMENT DATA**: Contains historical data of invoices whose payments are completed. To be used for building classification model to predict payment delays.
- **OPEN INVOICE DATA**: Contains open invoices. Need to use the classification model to assign payment delay probability to Vendors of the company for remedial measures.
- Checked data types and corrected them as needed.
- Dropped irrelevant columns such as 'CLASS' and handled missing values.
- Created new meaningful features like 'PAYMENT_TERM_DAYS,' 'RECEIPT_DAYS,' and 'PAYMENT_DAYS' from date columns.
- Cleaned the data by removing outliers and zero/negative USD amounts.

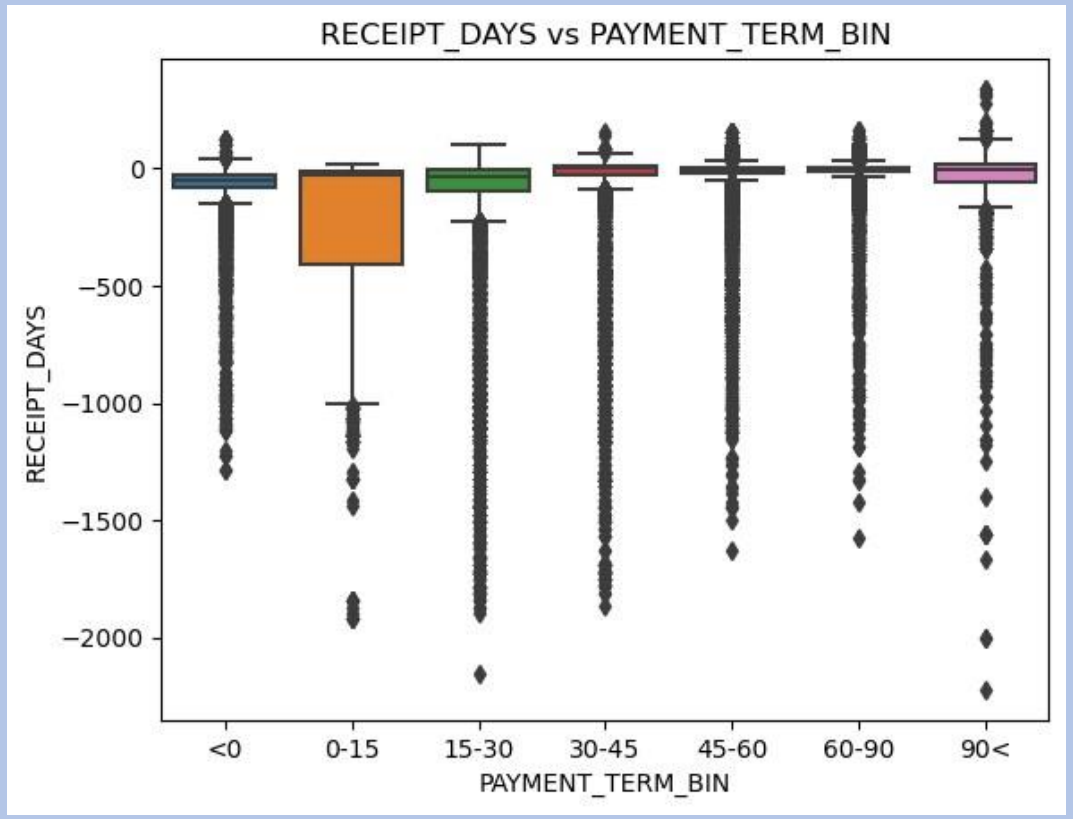
ANALYSIS APPROACH

Exploratory Data Analysis (EDA):

- Conducted univariate analysis of categorical variables.
- Reduced the number of categories in some columns by clubbing categories with lesser frequency.
- Analyzed the relationship between 'PAYMENT_TERM' and 'RECEIPT_DAYS'/ 'PAYMENT_DAYS' to identify payment delay patterns.

BI-VARIATE ANALYSIS



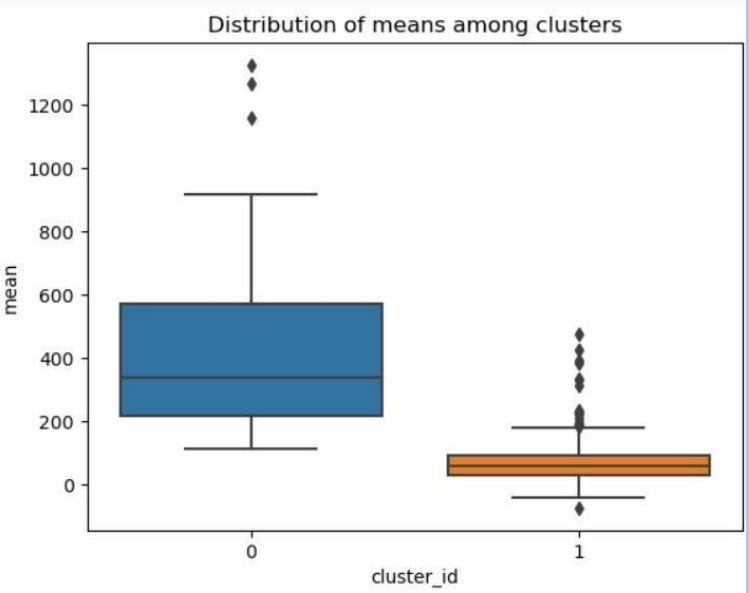


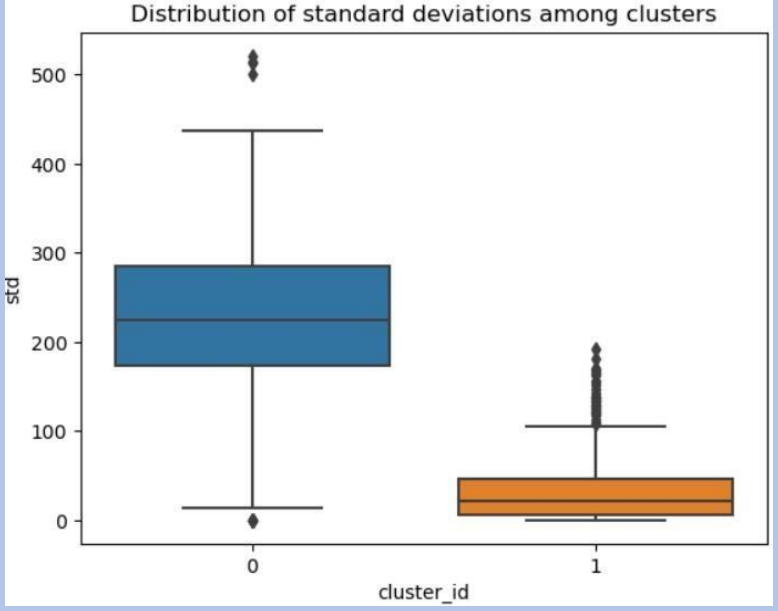
Among all the plots, the relation between PAYMENT_TERM_BIN and RECEIPT_DAYS appears to be significantly indicated the quantum of delay in payments is higher in PAYMENT_TERMS 0-15 days. So, the company may look into improving this to reduce major chunk of delays

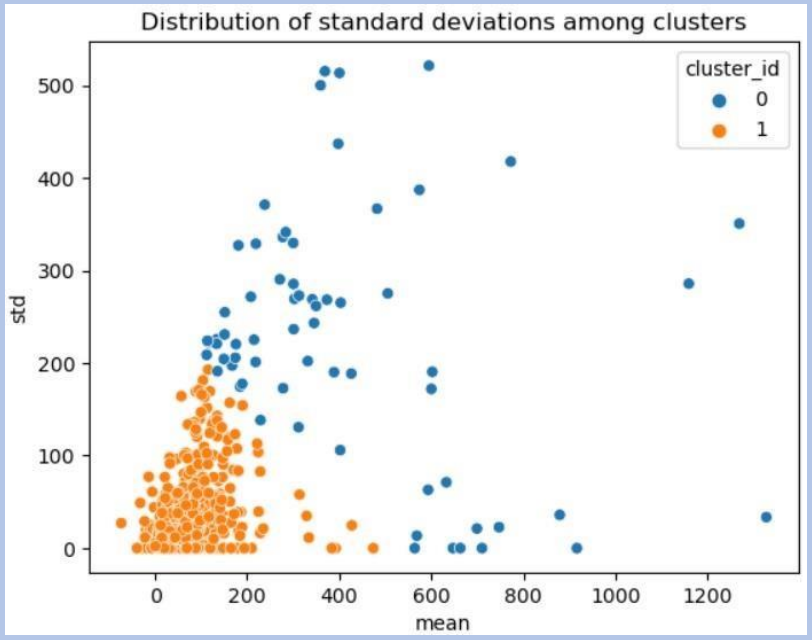
ANALYSIS APPROACH

Customer Segmentation:

- Created customer segments using clustering techniques based on payment behavior.
- Identified two customer segments with distinct payment patterns.
  - These ‘cluster _id’ are added to the data to be used as an independent feature in ML stage.







ANALYSIS APPROACH

Feature Engineering:

- Engineered additional features like 'Amount_ratio' , 'PAYMENT_TERM_BIN‘ and Dummies to enhance predictive power.

Model Building and Evaluation:

- Formulated the problem as a classification task to predict payment delays.
- Built classification models, including Logistic Regression, Decision Tree, and Random Forest. Along with Hyperparameter tuning.
- Evaluated model performance using ROC-AUC scores and selected DecisionTreeClassifier.

Interpretation:

- Interpreted the feature importance and identified critical factors for predicting payment delays, such as 'USD Amount' and 'PAYMENT_TERM.’

Model Building

Sl No	**ACCURACY**	**ROC_AUC**	**MODEL**
1	0.750617	0.685755	Logistic Regression
2	0.752176	0.663342	Base Decision Tree
3	0.860221	0.837149	Best Decision Tree
4	0.745466	0.673698	Base Random Forest
5	0.857319	0.819435	Best Random Forest

Based on above evaluation metrics,

Decision Tree Classifier with Hyperparameter tuning was selected because of highest ROC_AUC score

FEATURES AND THEIR COEFFICIENTS OF LOGISTIC REGRESSION CLASSIFIER

- Customers having PAYMENT_TERM <0 will incease probability of delay payments
- Customers having PAYMENT_TERM between 0-15 have have second highest effect on probablity of delay payments
- Customers having PAYMENT_TERM between 30-45,60-90, and INVOICE_CLASS_INV reduces likelyhood of delay payments
- As the ratio of Amount_ratio increases, customers are less likely to delay payments. It could mean vendors in places having USD as local currency, which would be mostly in USA would have lower probability of delays. This could be due to bottlenecks in Wiring money from different currencies.-
- It appears Invoices having SAR as the INVOICE_CURRECY_CODE would slightly increase chance of delaying payments

	Features	Coefficients
8	INVOICE_CLASS_INV	-1.403355
14	PAYMENT_TERM_BIN_60-90	-1.021347
12	PAYMENT_TERM_BIN_30-45	-0.825624
1	Amount_ratio	-0.799283
4	INVOICE_CURRENCY_CODE_SAR	0.857663
11	PAYMENT_TERM_BIN_15-30	1.066587
10	PAYMENT_TERM_BIN_0-15	1.722183
9	PAYMENT_TERM_BIN_<0	2.050505

	Feature	Importance
0	USD Amount log	0.522158
14	PAYMENT_TERM_BIN_60-90	0.110487
4	INVOICE_CURRENCY_CODE_SAR	0.074012
12	PAYMENT_TERM_BIN_30-45	0.073739
1	Amount_ratio	0.055949
2	INVOICE_CURRENCY_CODE_AED	0.031984
9	PAYMENT_TERM_BIN_<0	0.031494
17	cluster_id_1	0.025783
16	cluster_id_0	0.021598
11	PAYMENT_TERM_BIN_15-30	0.018689
5	INVOICE_CURRENCY_CODE_USD	0.011749
10	PAYMENT_TERM_BIN_0-15	0.005298
13	PAYMENT_TERM_BIN_45-60	0.004509
3	INVOICE_CURRENCY_CODE_Others	0.004363
7	INVOICE_CLASS_DM	0.002734
15	PAYMENT_TERM_BIN_90<	0.002718
8	INVOICE_CLASS_INV	0.002453
6	INVOICE_CLASS_CM	0.000284

FEATURES IMPORTANCE AS PER OF DECISION TREE CLASSIFIER

In the predictive model for payment delay, feature importance analysis reveals that "USD Amount log" plays the most significant role, followed by "PAYMENT_TERM_BIN_60-90" and "Amount_ratio." These features contribute the most to predicting whether customers are likely to delay payments, providing valuable insights for prioritizing collections efforts.

Top features in order of importance in the Decision Tree Classifier indicates similar insights to the one available in Logistic Regression Classifier.

ANALYSIS APPROACH

Real-world Prediction:

- Prepared open invoice data for predicting payment delays in practical scenarios.
- Preprocessed the data for modeling, considering only invoices with negative 'AGE' (yet-to-due date).
- Applied the trained model to predict payment delays at invoices level along with probabilties. These probabilities are aggregate at Customer level for scoring Customer probability for identifying customers who are likely to delay payment for follow-up action.

Threshold Setting and Aggregation:

- Set a threshold for payment delay probability based on business needs and available resources.
- Aggregated predictions to determine which customers are likely to delay payments.

ANALYSIS APPROACH

Customer Level Aggregation:

- A threshold can be set for payment delay probability based on business needs and available resources for regular follow-up with customers.

- Aggregated predictions probabilities can be used to determine which customers are likely to delay payments.

FEATURES AND THEIR COEFFICIENTS OF LOGISTIC REGRESSION CLASSIFIER

- **Customers having PAYMENT_TERM <0 will incease probability of delay payments**
- Customers having PAYMENT_TERM between 0-15 have have second highest effect on probablity of delay payments
- Customers having PAYMENT_TERM between 30-45,60-90, and INVOICE_CLASS_INV reduces likelyhood of delay payments
- As the ratio of Amount_ratio increases, customers are less likely to delay payments. It could mean vendors in places having USD as local currency, which would be mostly in USA would have lower probability of delays. This could be due to bottlenecks in Wiring money from different currencies.-
- It appears Invoices having SAR as the INVOICE_CURRECY_CODE would slightly increase chance of delaying payments

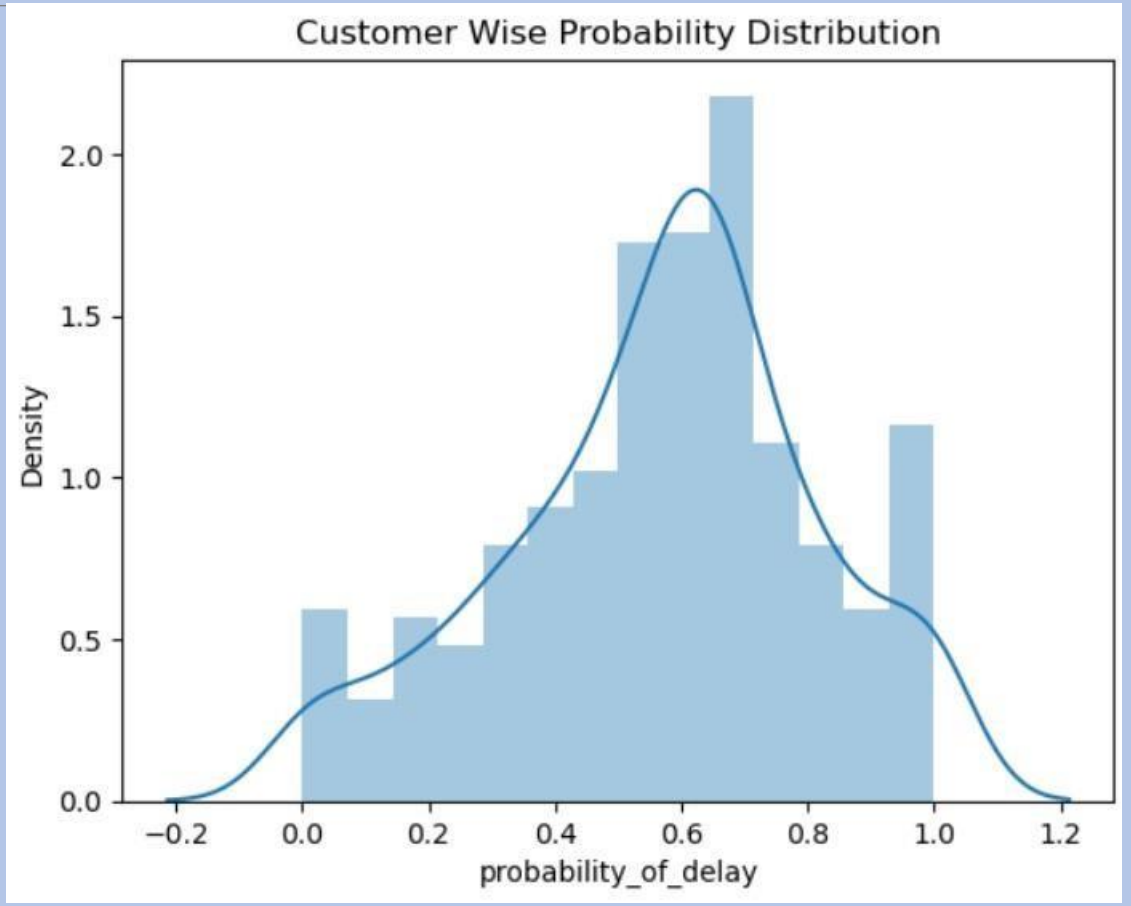
	Features	Coefficients
8	INVOICE_CLASS_INV	-1.403355
14	PAYMENT_TERM_BIN_60-90	-1.021347
12	PAYMENT_TERM_BIN_30-45	-0.825624
1	Amount_ratio	-0.799283
4	INVOICE_CURRENCY_CODE_SAR	0.857663
11	PAYMENT_TERM_BIN_15-30	1.066587
10	PAYMENT_TERM_BIN_0-15	1.722183
9	PAYMENT_TERM_BIN_<0	2.050505

Results

- We have plotted a boxplot for probabilities of customers to delay payment and cluster_id,

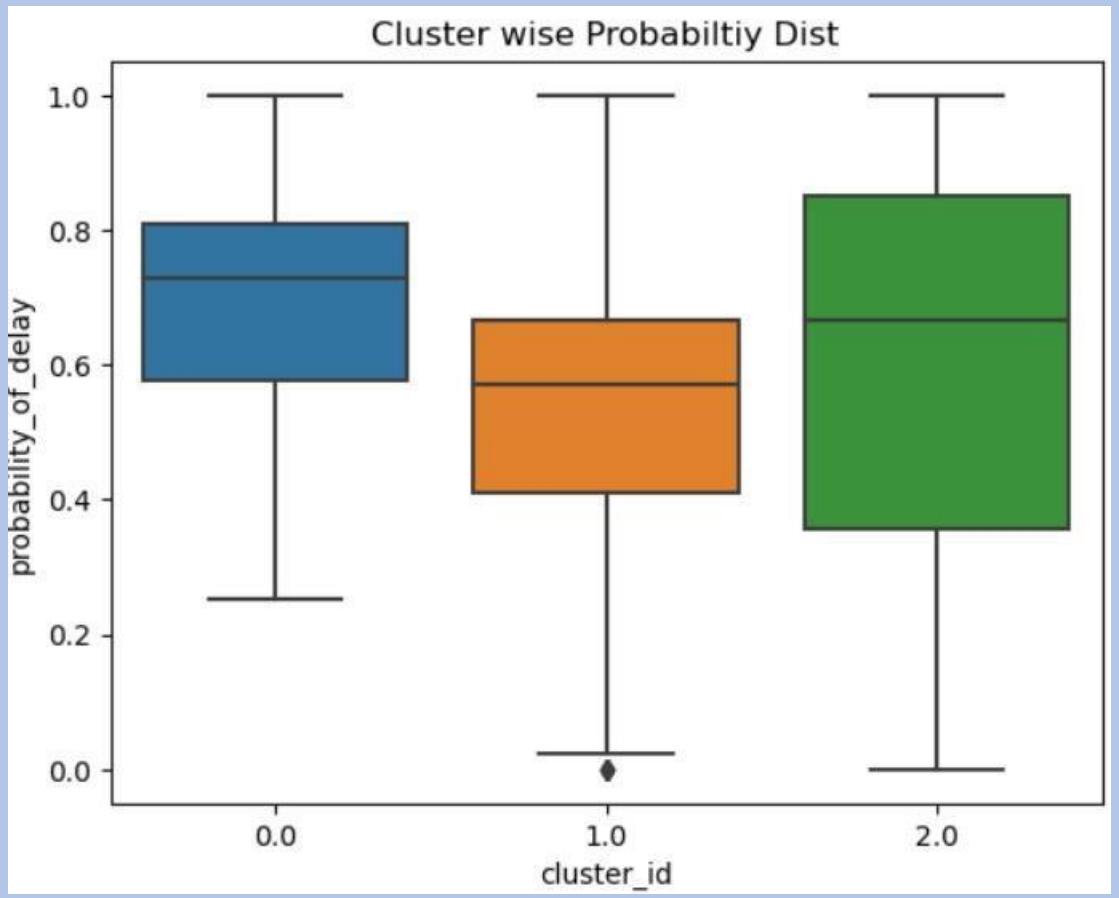
  Pred_data has the classification predictions of delayed payments at Invoice level along with invoice level delay probabilities

- This data can be used to aggregate at customer level to find probability of a customer to delay payment



Observations and Insights

- Cluster_id 2 in above plot was created to indicate Customer's whose customer_segment is not available due to lack of historical data
- Cluster_id 0, Customers had higher mean PAYMENT_DAYS with high standard deviation, has higher median probability for delayed payment
- Cluster_id 1,Custoemrs who had lower mean PAYMENT_DAYS with lower standard deviation has comparatively lower median of probability for delayed payments



OBSERVATIONS

Cluster_id 0 (High Mean PAYMENT_DAYS, High Standard Deviation):

- This customer segment, characterized by a higher mean PAYMENT_DAYS and greater variability in payment behavior, exhibits a higher median probability of delayed payments.
- The higher mean PAYMENT_DAYS suggests that, on average, this group tends to make payments later than other segments.
- The higher standard deviation indicates a wider spread of payment behavior within this segment, which may include both consistent late payers and occasional late payers.

Conclusion: Cluster_id 0 represents customers with a higher likelihood of delayed payments, and Schuster may need to pay extra attention to this segment to ensure timely collections.

Cluster_id 1 (Low Mean PAYMENT_DAYS, Low Standard Deviation):

- This customer segment, characterized by a lower mean PAYMENT_DAYS and lower variability in payment behavior, shows a comparatively lower median probability of delayed payments.
- The lower mean PAYMENT_DAYS suggests that, on average, this group tends to make payments closer to the due date or even earlier.
- The lower standard deviation indicates a more consistent payment behavior within this segment, with fewer instances of delayed payments.

Conclusion: Cluster_id 1 represents customers with a lower likelihood of delayed payments. While this segment appears to be more reliable in terms of timely payments, it is still essential to monitor and maintain these relationships**.**

Based on the above distribution, 0.6 can be suggested as the threshold for the Payments follow up team to decide on which vendors to target for regular follow-up.

ACTION POINTS

- The company can run the new real world data through the classifier and get the probabilities aggregated at customer level to get list of customers who are likely to delay payment.

- Based on the above distribution, 0.6 can be suggested as the threshold for the Payments follow up team to decide on which vendors to target for regular followup. This can be varied depending on the business considerations of manpower available to pursue Customers since setting the threshold low would increase the number of customers to follow up.

- Customer names who are likely to delay are put in csv file as output called CUSTOMER_LIKELY_TODELAY.

-Company may reconsider whether 15 day payment term is actually possible, since invoices with 15 day payment term has higher likely hood for default.

-Try to gather payment history of all customers which would help in segmenting them