Problem Statement:

1.Schuster is a multinational retail company dealing in sports goods and accessories. Schuster conducts significant business with hundreds of its vendors, with whom it has credit arrangements. Unfortunately, not all vendors respect credit terms and some of them tend to make payments late. Schuster levies heavy late payment fees, although this procedure is not beneficial to either party in a long-term business relationship. The company has some employees who keep chasing vendors to get the payment on time; this procedure nevertheless also results in non-value-added activities, loss of time and financial impact. Schuster would thus try to understand its customers' payment behaviour and predict the likelihood of late payments against open invoices.

2.To understand how to approach this problem using data science, let's first understand the payment process at Schuster now. Every time a transaction of goods takes place with a vendor, the accounting team raises an invoice and shares it with the vendor. This invoice contains the details of the goods, the invoice value, the creation date and the payment due date based on the credit terms as per the contract. Business with these vendors occurs quite frequently. Hence, there are always multiple invoices associated with each vendor at any given time.

Goal:

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.

Data Understanding:

RECEIPT_METHOD In which method payments have been made

CUSTOMER_NAME Name of the customer/vendor

CUSTOMER_NUMBER Customer's unique identity number

RECEIPT_DOC_NO Reference number of the payment receipt

RECEIPT_DATE The date in which the payment has been made

CLASS As the payment against these invoices have already been received so

Transaction Class as PMT (short for Payment) assigned

CURRENCY_CODE Currency used for the payment

Local Amount Invoice value in local currency

USD Amount Invoice Value converted to USD

INVOICE_ALLOCATED Invoice number that has been allocated to a particular vendor

INVOICE_CREATION_DATE The date on which the invoice was created

DUE_DATE The date by which the payment was to be made

PAYMENT_TERM Days given to the vendor/customer for making the payments

INVOICE_CLASS Three types of Invoice classes - Credit Memo or Credit Note (CM), Debit Memo or Debit Note (DM) or Invoice (INV)

INVOICE_CURRENCY_CODE Currency code as per the invoice generated

INVOICE_TYPE Invoice created for physical goods or services (non-goods)

Finally, the target variable will be derived based on the suggested information "You need to derive it by checking whether the payment receipt date falls within, or after the due date. By doing so, you can create your binary target variable as 1 or 0."

Recommendation for Customer Segmentation using K-Means Clustering:

Customer-level attributes could also be important independent variables to be included in the model.

A customer-level attribute can be determined via customer segmentation. You have to segment your customers based on two derived variables: the average payment time in days for a customer and the standard deviation for the payment time.

Using clustering techniques would result in a few distinct clusters of customers, which can be used as an input variable for the ML model.

Final model:

SMOTE+TOMEK Combining Oversampling and Under-sampling

1. Tomek links can be used as an under-sampling method or as a data cleaning method.
2. Tomek links to the over-sampled training set as a data cleaning method. Thus, instead of removing only the majority class examples that from Tomek links, examples from both classes are removed.

Accuracy: 0.98

F1 score: 0.98

Recall: 0.98

Precision: 0.99

Classification report:

	precision	recall	f1-score	support
0	0.64	1.00	0.78	1092
1	1.00	0.98	0.99	27090
accuracy			0.98	28182
macro avg	0.82	0.99	0.88	28182
weighted avg	0.99	0.98	0.98	28182

Confusion matrix:

[[1088 4]

[619 26471]]

Output Summary Table:

	Cust id	actual	predicted	is_delayed
67288	34647	1	1	yes
60971	7530	1	1	yes
53170	7588	1	0	no
39162	45720	0	0	no
15138	2624	1	0	no
20187	20844	1	1	yes
59331	3997	0	1	yes
30267	34876	1	0	no
37858	45720	1	0	no
7244	3927	1	0	no

Top 10 influential factors

age

payment_term_50% advance payment and 50% upon receiving the shipment

payment_term_eom

payment_term_lcsight

payment_term_on consignment

invoice_currency_code_eur

invoice_currency_code_gbp

invoice_currency_code_kwd

invoice_currency_code_qar

invoice_type_non goods

Business Recommendations:

1. We should focus more on the time difference between Due Date and the Invoice Payment Date

2. Payment terms: 50% advance payment and 50% upon receiving the shipment, eom, lcsight and on-consignment variables need to be considered with greater attention.

3. Where the invoice currency codes are eur, gbp, kwd and qar, the risk is higher of delay payment.

4. Invoice type non-goods has a lower impact than Goods invoice type in delayed payment