Credit Score Classification

Agenda

Overview of the Dataset

Exploratory Data Analysis

Data Preprocessing

Model Development and Results

Problem statement

Problem Statement

A Global Finance Company wants to build a method to classify customer Credit Scores based on bank details and credit-related data of customers

Action Item

Based on customer credit information, develop a machine learning model to classify credit ratings for Customers

Dataset

- Contains customer identifiers, demographics, and financial data (income, accounts, SSN).
- Tracks credit behavior through loan details, payment history, and utilization metrics.

Overview of the Dataset | Feature descriptions

No.	Feature	Description
1	id	Unique identifier of each observation
2	customer_id	Customer identification code, allowing you to link multiple records to the same individual.
3	month	Timestamp when the record was saved, indicating when the data was collected
4	name	Customer name, which can be used for identification purposes.
5	age	Customer's age
6	ssn	Customer's Social Security Number (SSN), a unique identification number used for verification.
7	occupation	Customer's occupation or profession, which can help understand their employment status.
8	annual_income	Customer's annual income
9	monthly_inhand_salary	Monthly salary or income available to the customer after deductions.
10	num_bank_accounts	Number of bank accounts held by the customer, indicating their banking activity.
11	num_credit_card	Number of credit cards held by the customer, reflecting their credit usage.
12	interest_rate	Interest rates related to the customer's financial products, such as loans or credit cards.
13	num_of_loan	Number of loans the customer has, providing insight into their debt obligations.
14	type_of_loan	Types of loans the customer holds, which may include mortgages, personal loans, etc.
15	delay_from_due_date	Late payments since the due date for loans or credit cards
16	num_of_delayed_payment	Number of times the customer has made late payments.
17	changed_credit_limit	Changes in the customer's credit limit that may affect their credit usage.
18	num_credit_iniquiries	Number of credit inquiries made by the customer, potentially affecting their credit score.
19	credit_mix	Composition of the customer's credit accounts that may affect their credit profile.
20	oustanding_debt	Amount of outstanding debt of the customer.
21	credit_utilization_ratio	Ratio of credit used to total available credit, a critical factor in credit scoring.
22	credit_history_age	Age of the customer's credit history, affecting their credit credibility.
23	payment_of_min_amount	How the customer handles minimum payment amounts on credit cards or loans.
24	total_bmi_per_month	Total monthly Equated Monthly Installment (EMI) payments made by the customer.
25	amount_invested_monthly	Amount the customer invests monthly, if applicable.
26	payment_behaviour	Customer behavior related to their payments, reflecting their financial responsibility.
27	monthly_balance	Monthly balance in the customer's financial accounts.
28	credit_score	Target variable representing the customer's credit score

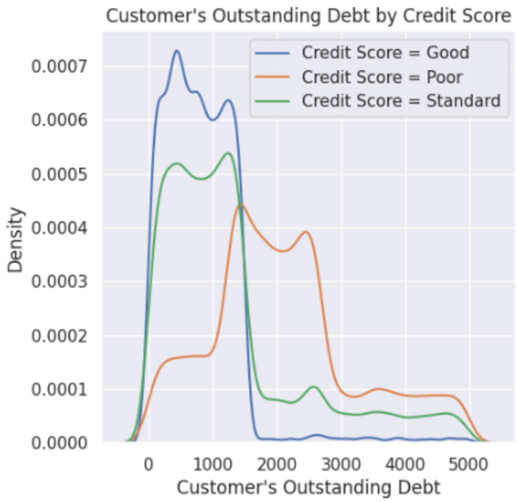
Target

Agenda

- Overview of the Dataset
- Exploratory Data Analysis
- Data Preprocessing
- Model Development and Results

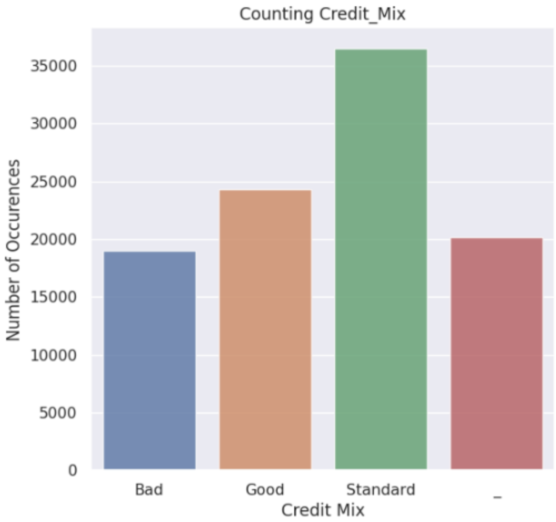
Features Analysis

Customers in the Good Credit Score group typically have smaller debt amounts compared to customers in the Standard and Poor Credit Score groups

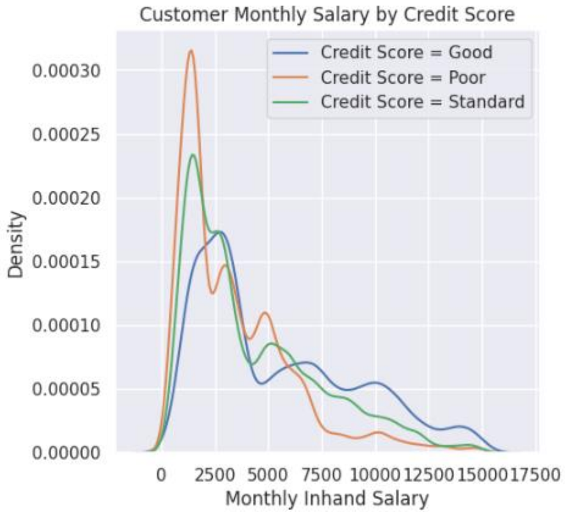


The majority of customers classified with 'Good' credit have not made minimum payments on their loans. Conversely, customers with 'Poor' Credit Scores typically make minimum payments on their loans

The majority of customers are classified with a Standard Credit_Mix



Monthly salary of customers in the Poor Credit Score group is typically lower than those in the Standard and Good Credit Score groups



Agenda

- Overview of the Dataset
- Exploratory Data Analysis
- Data Preprocessing
- Model Development and Results

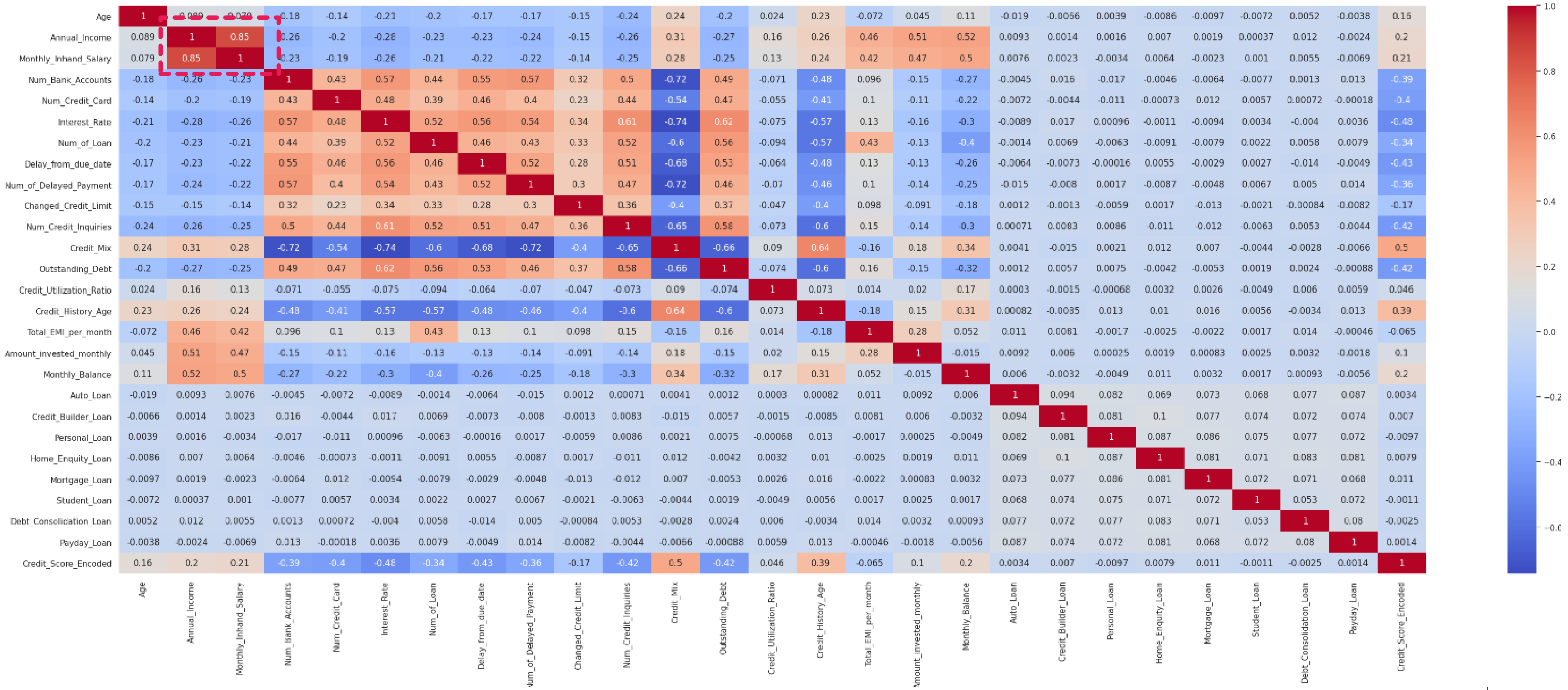
Data Preprocessing | Problems and Solutions

Problems	Details	Solutions
Inappropriate data types	• Data type of some columns needs to be converted from Object to Category • Data type of some features needs to be converted to float	• Convert data types of variables Month, Occupation, Type_of_Loan, Credit_Mix, Payment_of_Min_Amount, Payment_Behaviour, Credit_Score from object to Category • Remove special characters (such as _) from columns that need to be converted to float or int data types
Missing values	• Columns Monthly_Inhand_Salary, Type_of_Loan, Name, Credit_History_Age, Num_of_Delayed_Payment, ... have missing values	• Missing values will be replaced using one of the following two methods based on the feature's data type: • Replace missing values with the median of that feature • Replace missing values with the mode of that feature based on unique Customer_ID
Values in some columns do not align with business logic	• Values in the Num_Bank_Accounts column must be greater than 0 • Values in the Num_of_Loans and Delay_from_due_date columns must be greater than or equal to 0	• Update values to ensure they are correct according to business requirements
Encoding data from text to numbers	• Encoding for the following feature columns: Credit_Mix, Payment_of_Min_Amount, Occupation and the target column Credit_Score	• Encode categorical data to numerical form using sklearn

Data Preprocessing | Problems and Solutions

Problems	Details	Solutions
The Type_of_loan column cannot be used for model training	• Most loan information is currently declared and mixed as String format	• Each loan type column needs to be split into multiple columns corresponding to each loan type. Assign a value of 1 or 0 for each new Type_of_Loan column
Most values in the Occupation column have the character '_____' as Job	• These values cannot be used for model training	• Will replace '_____' values with the mode based on unique Customer_ID
Standardization	• Balancing based on Data Distribution	• Will apply relevant techniques
Handling Outliers	• Most columns have outliers that need to be removed from the features to enable the model to produce more accurate results	• Replace outliers with Q1 or Q3

Data Preprocessing| Correlation Matrix



Removing the Monthly_Inhand_Salary feature due to high correlation with the Annual_Income feature (correlation coefficient of 0.85)

Agenda

- Overview of the Dataset
- Exploratory Data Analysis
- Data Preprocessing
- Model Development and Results

Model Development| Results

Summarize Key Insights

Results Log:

No	Model	Accuracy	Poor Credit Score (0)			Standard Credit Score (1)			Good Credit Score (2)			Confusion Matrix
			Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	
1	Logistic Regression	48%	58%	74%	65%	67%	26%	38%	31%	72%	43%	[[17163 2158 3820] [11704 11153 19680] [651 3294 10377]]
2	Gaussian Naives Bayes	56%	61%	74%	67%	83%	36%	50%	37%	87%	52%	[[17142 1666 4333] [10669 15160 16708] [449 1473 12400]]
3	Decision Tree	66%	63%	65%	64%	71%	69%	70%	55%	58%	57%	[[15060 6955 1126] [7649 29299 5589] [1050 4934 8338]]
4	KNN (k= 3)	60%	59%	72%	65%	72%	52%	61%	44%	63%	51%	[[16772 4695 1674] [10362 22257 9918] [1406 3956 8960]]
5	KNN (k= 5)	58%	58%	73%	64%	71%	48%	57	41	63	50	[[16887 4311 1943] [11059 20272 11206] [1403 3873 9046]]
6	KNN (k= 7)	56%	58	72	64	71	45	55	38	63	48	[[16756 4048 2337] [11048 18959 12530] [1319 3863 9140]]
7	RF - 50 trees	75%	73	78	75	81	74	77	65	74	69	[[18059 3995 1087] [6459 31416 4662] [242 3487 10593]]
8	RF - 100 trees	75%	73	78	76	81	74	77	65	75	69	[[18148 3870 1123] [6384 31403 4750] [228 3396 10698]]
9	Bagging	72%	70	74	72	79	70	74	58	74	65	[[17027 4381 1733] [6980 29716 5841] [322 3369 10631]]
10	AdaBoostClassifier(n_estimators=100)	66%	63	66	65	78	63	69	51	75	61	[[15353 4715 3073] [8560 26693 7284] [533 3013 10776]]
11	GradientBoostingClassifier()	69	67	73	70	81	65	72	53	78	63	[[16897 3650 2594] [7917 27464 7156] [320 2813 11189]]
12	XGBClassifier	73%	72	73	72	77	74	76	62	69	65	[[16829 5066 1246] [6363 31491 4683] [331 4170 9821]]
13	LightGBM	72%	70	74	72	79	70	74	58	74	65	[[17027 4381 1733] [6980 29716 5841] [322 3369 10631]]

Conclusion

- Recommendation to use the Random Forest Model as it provides the best results based on Accuracy, Precision, and Recall for each Target Class
- Bagging and Boosting methods can also be used as their evaluation metrics are relatively high
- Finally, the company should consider adjusting the system to ensure information in the records is captured more completely