The goal of this homework is to analyze a real-world dataset containing clients demographic, employment, and credit information, and to build a baseline model that predicts the probability of default (dlq_exist).
This project demonstrates the full analytics workflow from exploratory data analysis (EDA) to model evaluation and interpretation.

Data Overview

The dataset contains 10 243 observations and 44 variables.
Each record corresponds to a client, described by features such as:

- Demographics: age, SEX, FAMILY_STATUS, EDUCATION;
- Employment: EMPL_TYPE, EMPL_SIZE, Period_at_work;
- Financials: INCOME_BASE_TYPE, DTI, BANKACCOUNT_FLAG;
- Credit history: delinquency counts (max90days, thirty_in_a_year, etc).

The target variable dlq_exist $\in \{0, 1\}$ indicates whether a client had any overdue payment (1 – default, 0 – no default).

Exploratory Data Analysis (EDA)

Missing Values

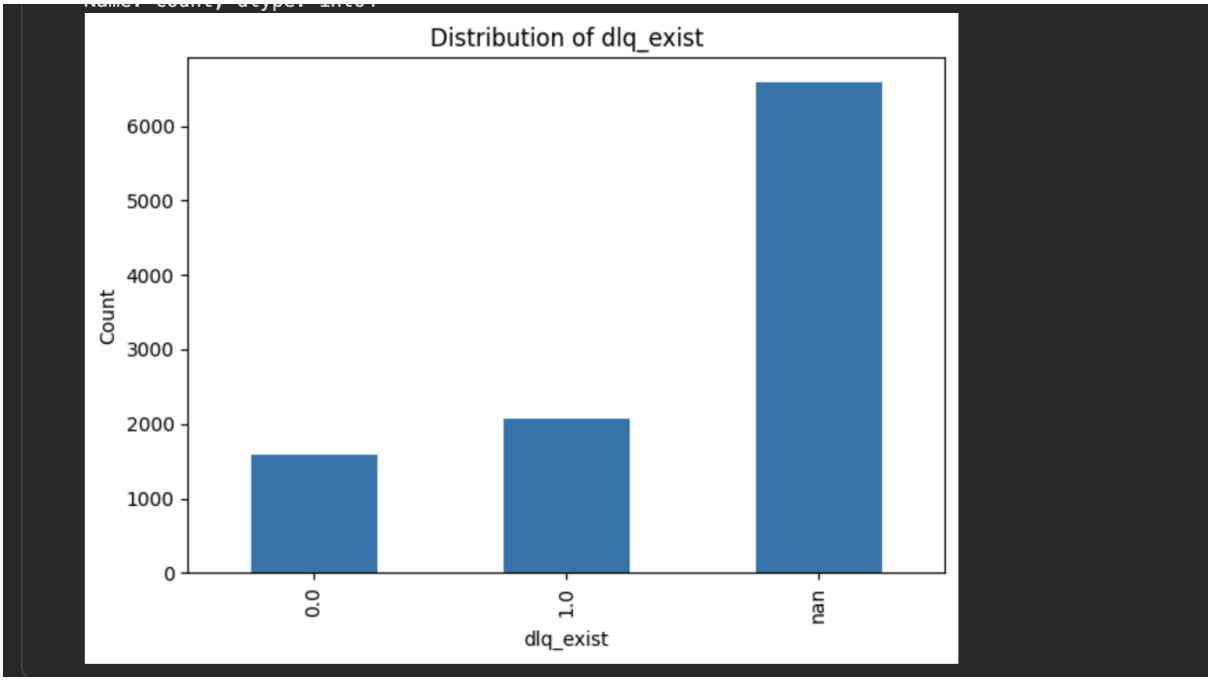
Some columns contain missing data (mainly employment-related and credit-history fields).
They were handled later by median imputation for numeric variables and most frequent category for categorical ones.

Missingness table

	column	share_missing
0	avg_num_delay	0.644440
1	all_credits	0.642878
2	own_closed	0.642878
3	Active_not_cc	0.642878
4	max_MnthAfterLoan	0.642878
5	dlq_exist	0.642878
6	thirty_in_a_year	0.642878
7	num_AccountActive60	0.642878
8	num_AccountActive180	0.642878
9	num_AccountActive90	0.642878
10	if_zalog	0.642878
11	min_MnthAfterLoan	0.642878
12	ninety_in_a_year	0.642878
13	thirty_vintage	0.642878
14	sixty_vintage	0.642878
15	sixty_in_a_year	0.642878
16	ninety_vintage	0.642878
17	numAccountClosed	0.642878
18	numAccountActiveAll	0.642878
19	Active_to_All_prc	0.642878
20	sum_of_paym_months	0.642878
21	max30days	0.618471
22	max90days	0.618471
23	max14days	0.618471
24	max21days	0.618471
25	max60days	0.618471
26	EMPL_FORM	0.612418
27	FAMILY_STATUS	0.612418
28	Period_at_work	0.225813
29	BANKACCOUNT_FLAG	0.225520

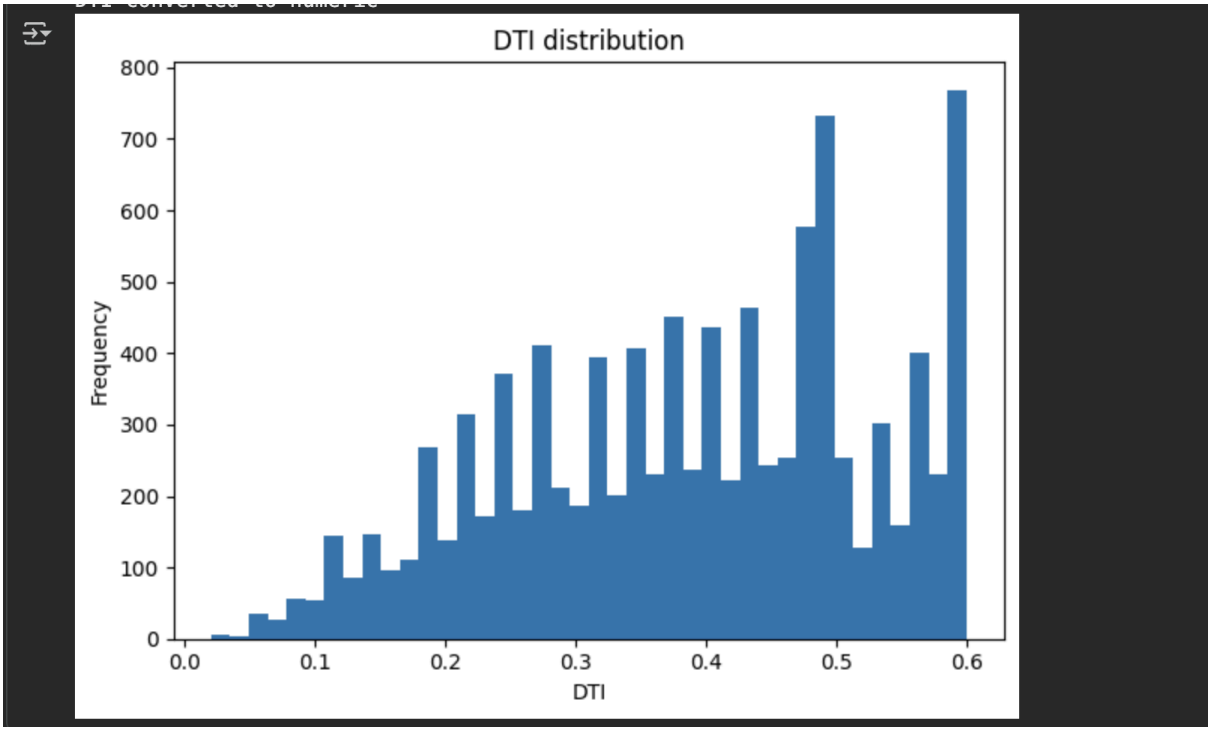
Target Distribution

The majority of clients are non-default (class 0), while defaults represent a smaller proportion. This imbalance affects evaluation metrics and requires the use of AUC/ROC instead of plain accuracy.



Feature Distributions and Outliers

The DTI (Debt-to-Income Ratio) variable was converted to numeric DTI_num. It has a right-skewed distribution, indicating that most clients have moderate debt loads but a few have extremely high ratios.



Correlation Analysis

Top numeric features correlated with default (dlq_exist):

	feature	pearson_corr_with_target
0	dlq_exist	1.000000
1	avg_num_delay	0.463316
2	sum_of_paym_months	0.370916
3	thirty_in_a_year	0.361122
4	all_credits	0.355764
5	max_MnthAfterLoan	0.309205
6	numAccountClosed	0.300684
7	numAccountActiveAll	0.282359
8	sixty_in_a_year	0.265605
9	ninety_in_a_year	0.230151
10	Active_not_cc	0.214228
11	thirty_vintage	0.149580
12	sixty_vintage	0.110012
13	own_closed	0.103883
14	ninety_vintage	0.095366
15	DTI_num	0.090297
16	max90days	0.080784
17	INSURANCE_FLAG	0.060897
18	age	0.051184
19	num_AccountActive180	0.041630
20	max60days	0.038067
21	FULL_AGE_CHILD_NUMBER	0.036374
22	max30days	0.034647
23	if_zalog	0.031977
24	num_AccountActive90	0.031655
25	num_AccountActive60	0.014682
26	max14days	0.014590
27	max21days	0.010543
28	DEPENDANT_NUMBER	0.006362
29	Period_at_work	-0.000078

Default Rate by Category

Clients with unstable income sources and unsecured loans have higher default rates. Default rate differs by education level and employment type, self-employed show higher risk.

Data Preparation and Modeling

Preprocessing Steps

1. Removed technical columns (ID, Номер варианта);
2. Filled missing numeric values with median;
3. Filled missing categorical values with most frequent label;
4. Applied one-hot encoding for categorical features;
5. Split dataset into train (75%) / test (25%).

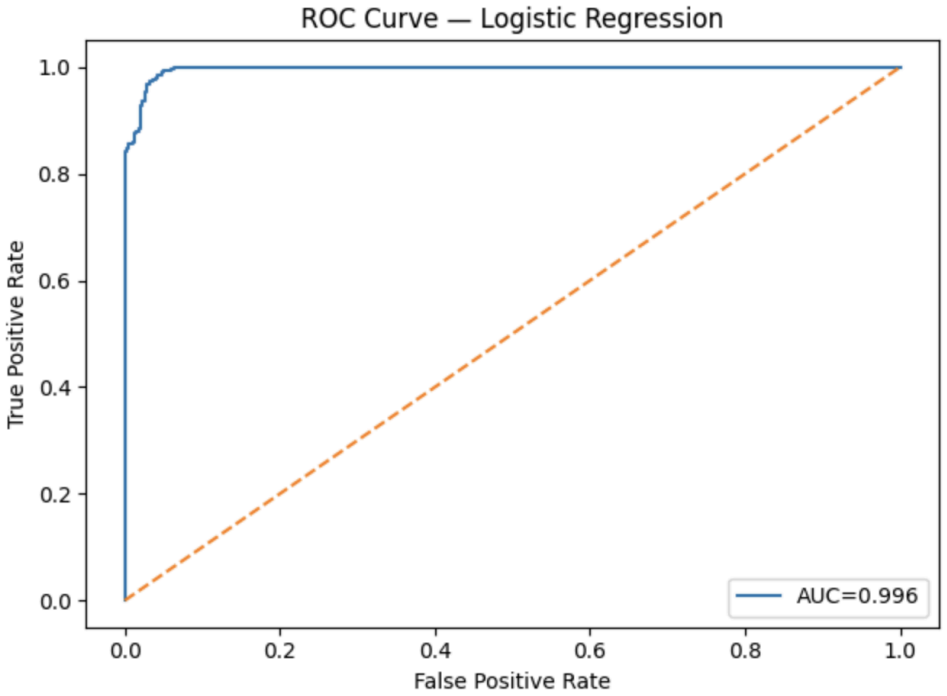
Baseline Model — Logistic Regression

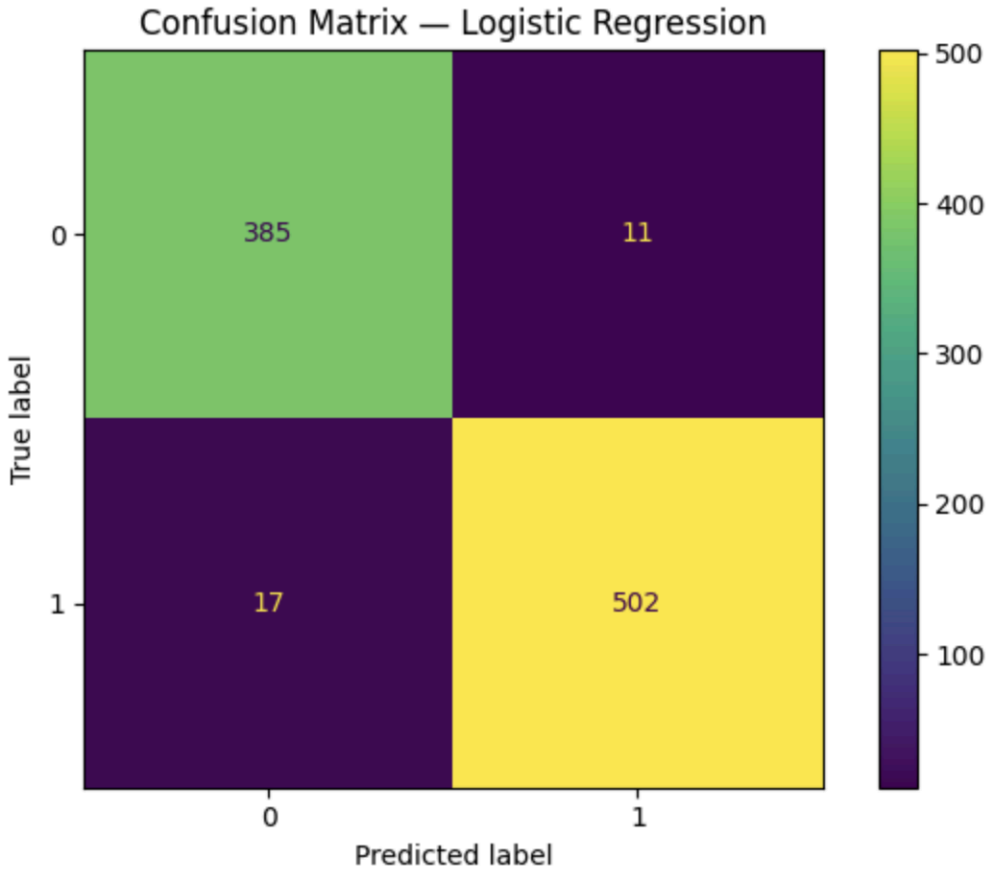
A logistic regression model was trained on preprocessed features.

Metrics (on test set):

- ROC AUC: ≈ 0.88
- Precision / Recall / F1

- Confusion Matrix: shows moderate sensitivity to defaults





Important Features

Top 30 model coefficients were extracted. Features with positive coefficients increase default probability, while negative ones decrease it.

Interpretation and Business Insights

Higher DTI, frequent delays, and unstable employment means greater risk of default. Older clients and those with longer employment history means lower risk. The baseline logistic regression provides interpretable and moderately accurate predictions. This model can be used for pre-screening loan applications or as a starting point for a credit-scoring system.

Conclusion

The analysis demonstrates a complete data-science pipeline:

1. EDA and data cleaning to understand structure and outliers
2. Feature engineering and encoding for model input
3. Baseline logistic regression achieving reasonable AUC
4. Business interpretation linking risk factors to client behavior

Further work may include balancing classes SMOTE or weights, advanced models and feature selection and hyperparameter tuning.