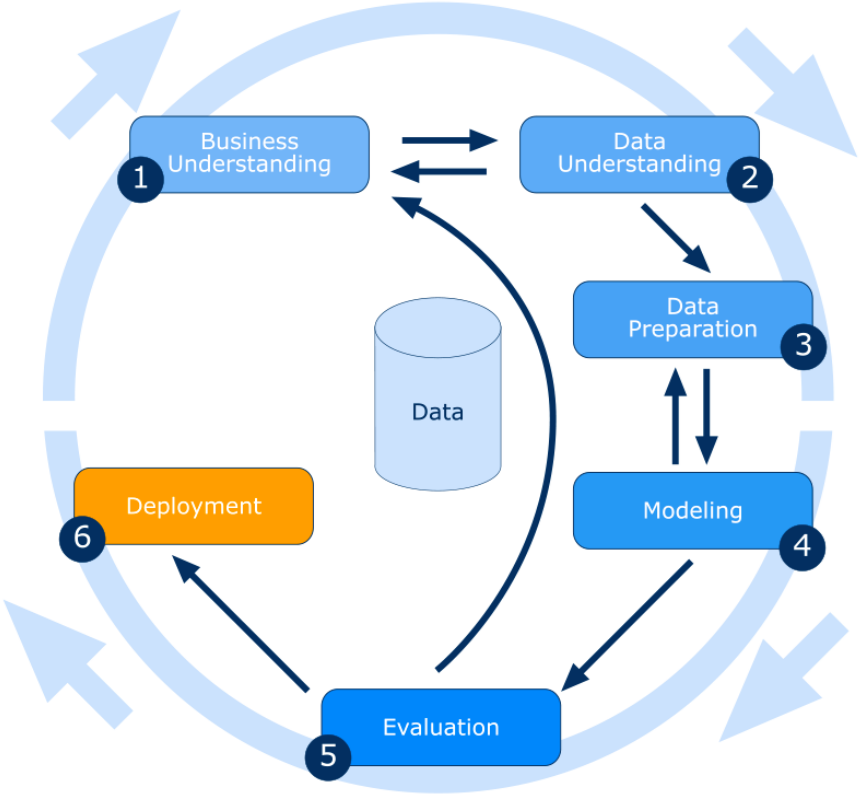
Addressing Inefficiencies in Traditional Credit Scoring Methods

Seth Abayo, Lanston Chen, Guangming (Dola) Qiu, Yizhou (Paul) Sun, Yihua (Anthony) Wang

Group X

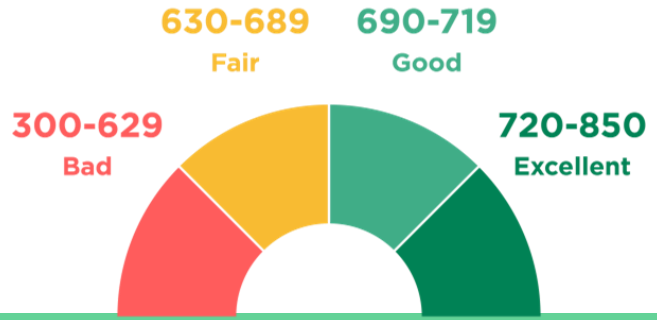
Data Mining Process: CRISP



Business Understanding - Credit Scoring

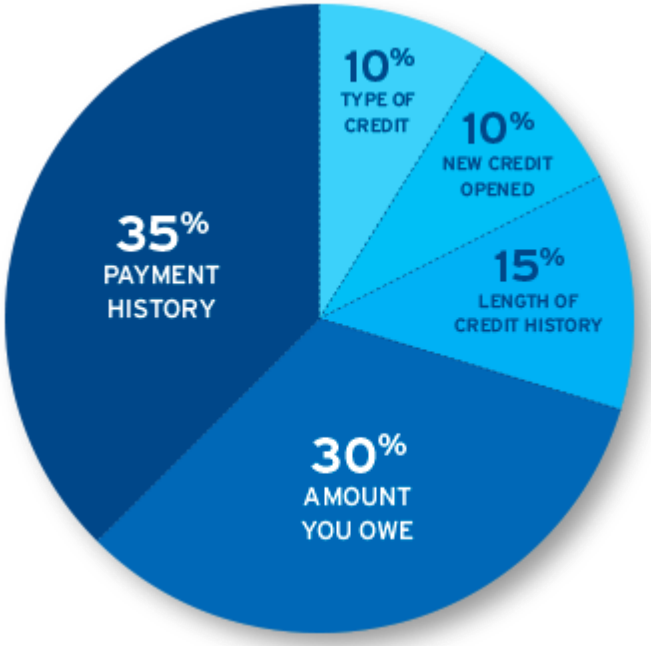
Credit scoring is a system used by lenders and financial institutions to assess the creditworthiness or risk of a potential borrower. It's a numerical representation of a person's credit behavior and financial history, and it's used to predict the likelihood that the person will repay borrowed money.

A **high credit score** indicates that the person has consistently paid their bills. The person seems to be less risky to lenders and therefore can get better loan terms.



Inefficiencies in traditional credit scoring methods

- **Lack of Comprehensive Data:** Traditional methods often rely on a limited set of financial factors, potentially overlooking other relevant indicators of an individual's creditworthiness.
- **Limited Historical Data:** New borrowers or those with limited credit histories ("thin files") might not have enough data points for a comprehensive evaluation, leading to inadequate or unfair scoring.
- **Time Delays:** Manual checks and verifications can slow down the loan approval process, leading to delays for consumers.



Data Understanding: Dataset

The Bank Credit-related Information Dataset

- Records 12,500 customers and their information with credit score
- Each customer has 8 rows set including from January to August
- Totally 100,000 rows of data
- Target Variable: Credit_Score

Data Understanding: Main Bank Clients Features

- **Income:**
  - Annual income
  - Monthly inhand salary
- **Account Information:**
  - Number of other credit cards owned by this customer
  - Interest rate on this credit card
  - Monthly balance amount from the customer
- **Loan Repayment:**
  - Average number of days delayed from the payment date
  - Whether the customer only pay the minimum amount of the loan
  - The fixed payment amount made by a borrower to a lender at a specified date each calendar month

Data Understanding: Example

ID	Customer_ID	Month	Name	Age	SSN	Occupation
0x1602	CUS_0xd40	January	Aaron Maashoh	23	821-00-0265	Scientist
Annual_Income	Monthly_Inhand_Salary	Num_Bank_Accounts	Num_Credit_Card	Interest_Rate	Num_of_Loan	Type_of_Loan
19114.12	1824.843333	3	4	3	4	Auto Loan, Credit-Builde
Delay_from_due_date	Num_of_Delayed_Paym	Changed_Credit_Limit	Num_Credit_Inquiries	Credit_Mix	Outstanding_Debt	Credit_History_Age
3	7	11.27	4	Good	809.98	22 Years and 1 Months
Payment_of_Min_Amount	Total_EMI_per_month	Amount_invested_monthly	Payment_Behaviour	Monthly_Balance	Credit_Score	
No	49.57494921	80.41529544	High_spent_Small_value_payments	312.4940887	Good	

Target Variable

Data Understanding: Exploratory Data Analysis

01 **Data Summary**

- Data Info
- Data description

02 **Checking Values**

- Unique Values for Categorical Variable
- Null Value
- Duplicate Rows

03 **Checking Distribution**

- For Predictors
- For Target Variable

Data Understanding: EDA – Summary & Value Check

```
RangeIndex: 100000 entries, 0 to 99999
Data columns (total 28 columns):
 #   Column                    Non-Null Count   Dtype
---  ------                    --------------   -----
 0   ID                        100000 non-null  object
 1   Customer_ID               100000 non-null  object
 2   Month                     100000 non-null  object
 3   Name                      90015 non-null   object
 4   Age                       100000 non-null  object
 5   SSN                       100000 non-null  object
 6   Occupation                100000 non-null  object
 7   Annual_Income             100000 non-null  object
 8   Monthly_Inhand_Salary     84998 non-null   float64
 9   Num_Bank_Accounts         100000 non-null  int64
 10  Num_Credit_Card           100000 non-null  int64
 11  Interest_Rate             100000 non-null  int64
 12  Num_of_Loan               100000 non-null  object
 13  Type_of_Loan              88592 non-null   object
 14  Delay_from_due_date       100000 non-null  int64
 15  Num_of_Delayed_Payment    92998 non-null   object
 16  Changed_Credit_Limit      100000 non-null  object
 17  Num_Credit_Inquiries      98035 non-null   float64
 18  Credit_Mix                100000 non-null  object
 19  Outstanding_Debt          100000 non-null  object
 20  Credit_Utilization_Ratio  100000 non-null  float64
 21  Credit_History_Age        90970 non-null   object
 22  Payment_of_Min_Amount     100000 non-null  object
 23  Total_EMI_per_month       100000 non-null  float64
 24  Amount_invested_monthly   95521 non-null   object
 25  Payment_Behaviour         100000 non-null  object
 26  Monthly_Balance           98800 non-null   object
 27  Credit_Score              100000 non-null  object
```

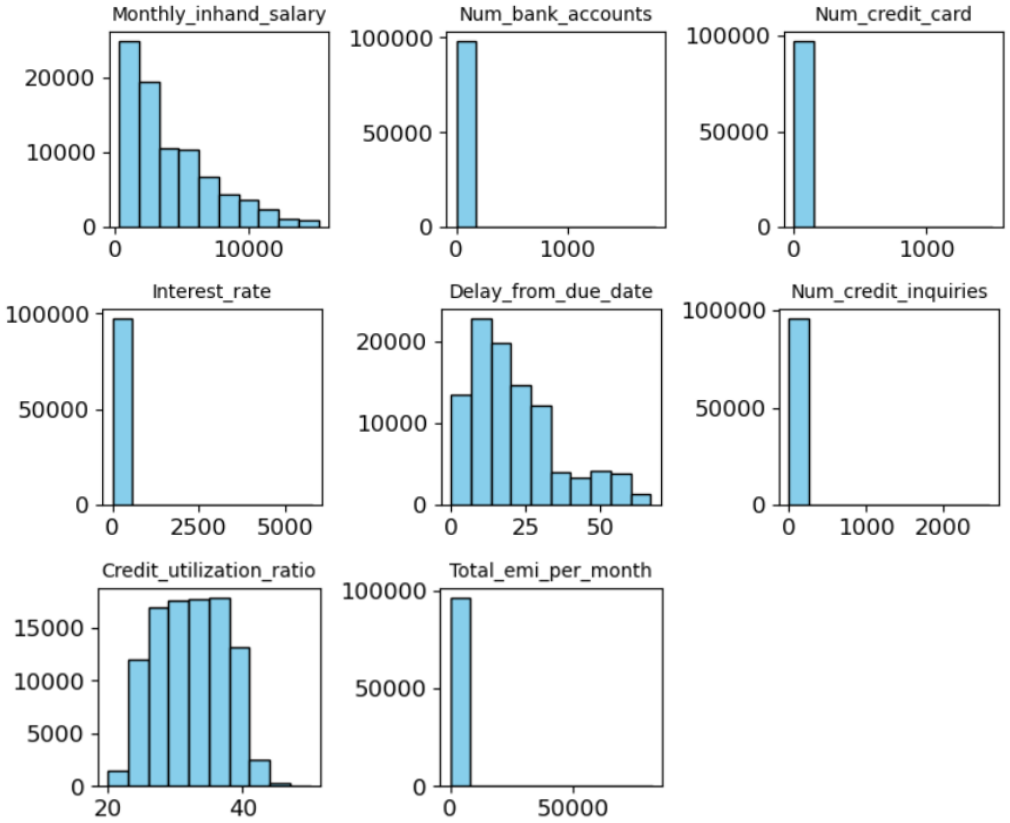
	Monthly_Inhand_Salary	Num_Bank_Accounts	Num_Credit_Card
count	84998.000000	100000.000000	100000.00000
mean	4194.170850	17.091280	22.47443
std	3183.686167	117.404834	129.05741
min	303.645417	-1.000000	0.00000
25%	1625.568229	3.000000	4.00000
50%	3093.745000	6.000000	5.00000
75%	5957.448333	7.000000	7.00000
max	15204.633333	1798.000000	1499.00000

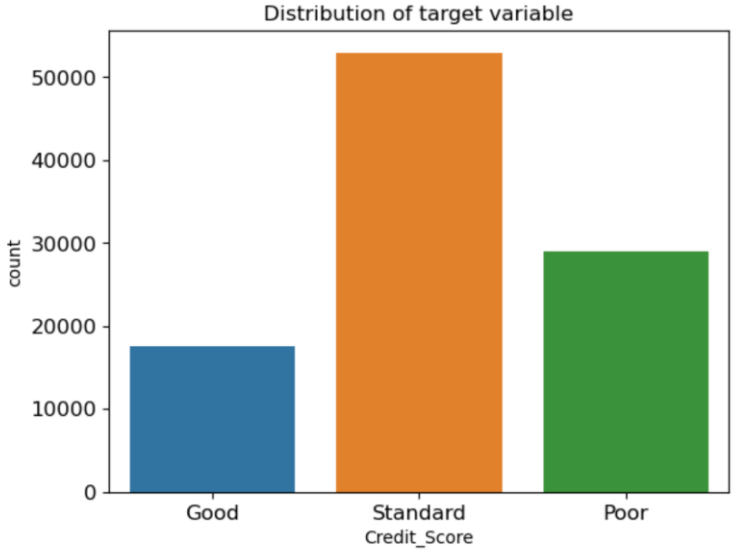
```
Data has duplicate lines: 0
```

```
array(['Good', 'Standard', 'Poor'], dtype=object)
```

```
ID                          0
Customer_ID                 0
Month                       0
Name                     9985
Age                         0
SSN                         0
Occupation                  0
Annual_Income               0
Monthly_Inhand_Salary   15002
Num_Bank_Accounts           0
Num_Credit_Card             0
Interest_Rate               0
Num_of_Loan                 0
Type_of_Loan            11408
Delay_from_due_date         0
Num_of_Delayed_Payment   7002
Changed_Credit_Limit        0
Num_Credit_Inquiries     1965
Credit_Mix                  0
Outstanding_Debt            0
Credit_Utilization_Ratio    0
Credit_History_Age       9030
Payment_of_Min_Amount       0
Total_EMI_per_month         0
Amount_invested_monthly  4479
Payment_Behaviour           0
Monthly_Balance          1200
Credit_Score                0
dtype: int64
```

Data Understanding: EDA – Checking Distribution





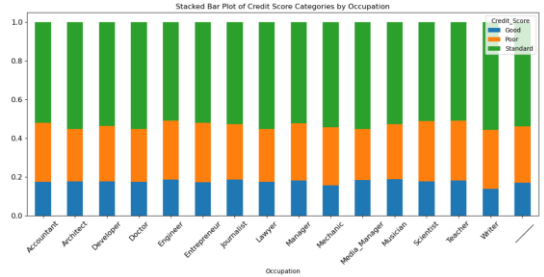
Data Preprocessing: Exclude Variables

Dropped categorical variables that has no relationship for Credit Score:

- ID
- Customer_ID
- Name
- SSN
- Month

Dropped nominal categorical variables which cannot make classification for Credit Score:

- Type_of_Loan
- Payment_Behaviour
- Occupation



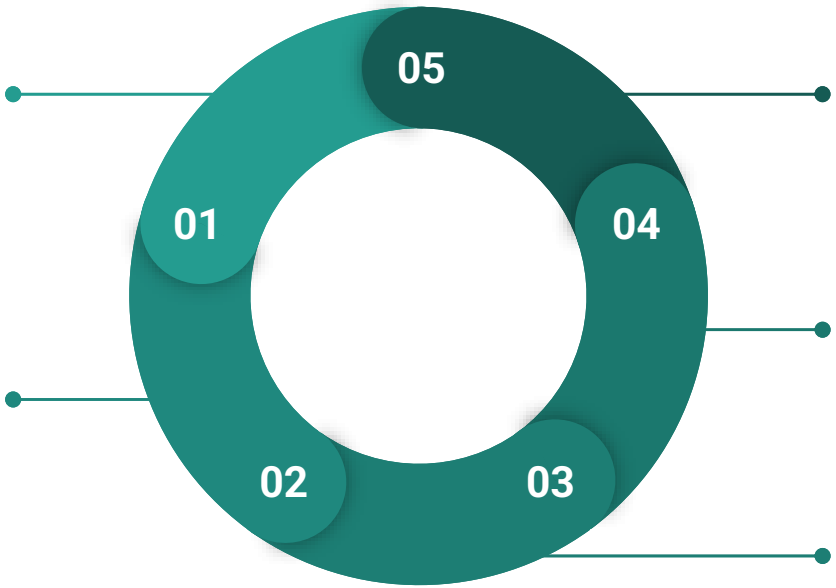
Dropped variables which would lead to Data Leakage:

- Credit_Mix
- Interest_Rate

Data Preprocessing: Cleaning Entries

Deleted special character like “_” to convert some categorical variable back to numerical

Changed the expression of entries like “22 years and 4 months” into “22.333” to make it numerical



Imputing all existing missing values to median of the column for numerical variables, mode for categorical

Used 10 IQR range from median to deleted unrealistic values

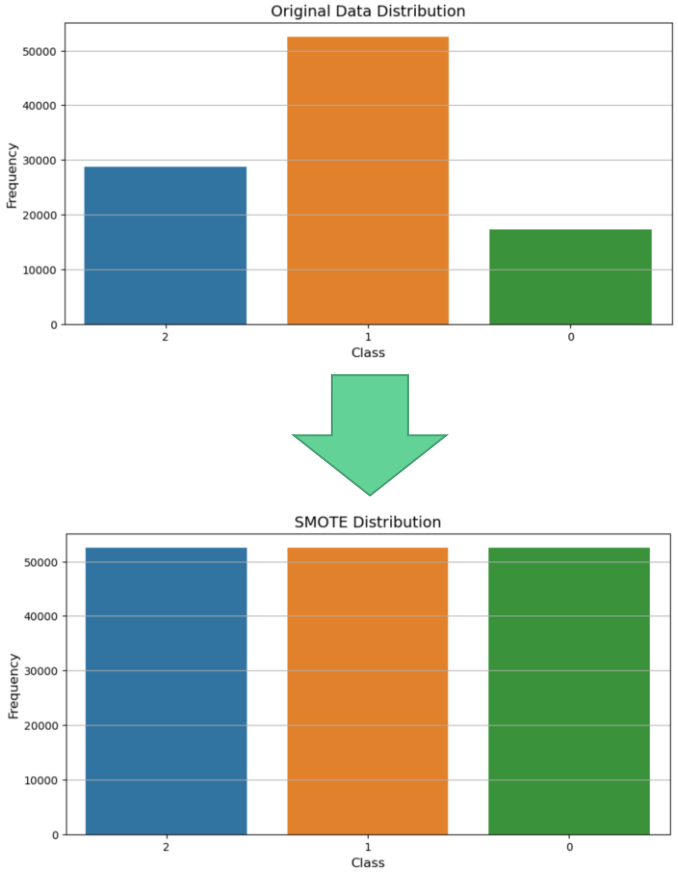
Limited variables in a reasonable range (e.g 18 < Age <80) base on empirical method to drop incorrect record

Data Preprocessing: Target Variable

Transformed the ordinal Target Variable: Good, Standard, Poor to numerical ranking as: 0, 1, 2

Used Synthetic Minority Over-sampling Technique (SMOTE) to balance Target Variable

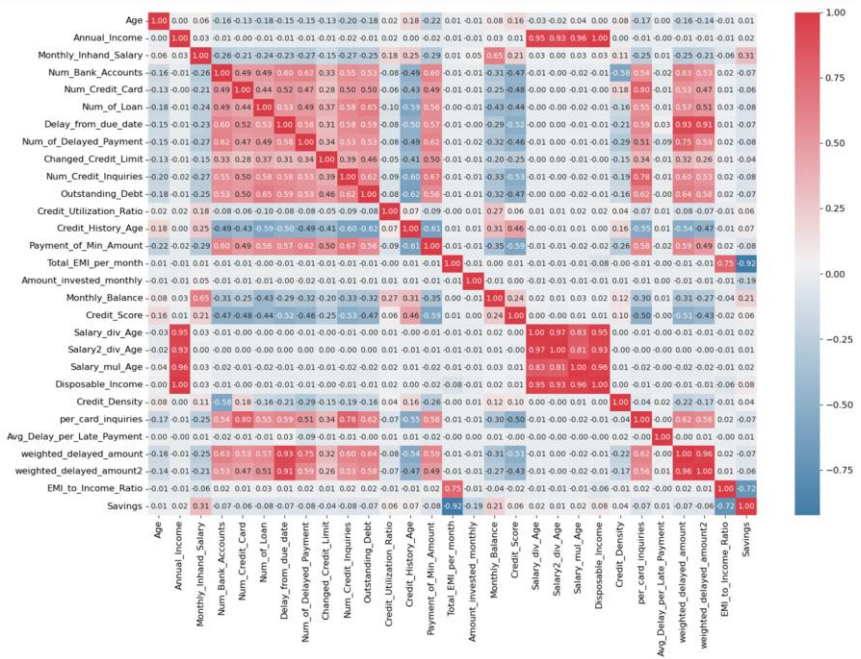
- **Mitigating Loss of Information**
- **Synthetic Sample Creation**
- **Improved Model Performance**



Data Preprocessing: Feature Engineering

Step 1: Features Creation → **Step 2:** Build Correlation Matrix → **Step 3:** Features Selection

Age Related
Income & EMI
Bank & Credit info
Payment Delays & Credit Behavior
Debt & Liabilities
Investment & Savings



Keep New Features with High Coefficient

Drop Features with Low Correlation to Avoid Overfitting

Modeling

- Experiment with various models to identify the top-performing one
  - – Logistic Regression
  - – K-NN
  - – Decision Tree
- Use nested cross-validation for hyper parameter tuning
  - 5 folds in the inner and outer loops
  - Choose F1 score as the metric
- Evaluate the generalization performance

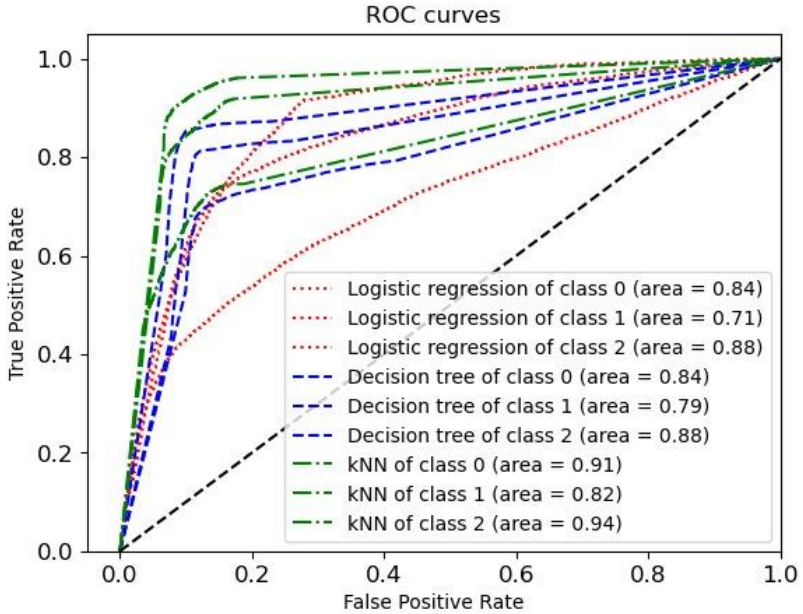
Modeling

Model	F1 Score	Accuracy	Kappa Score
Decision Tree	0.77	0.77	0.66
k-NN	0.75	0.76	0.64
Logistic Regression	0.45	0.50	0.25

The **Decision Tree** model achieved the highest F1 score, while **Logistic Regression** exhibited relatively low score. One possible explanation for this difference could be attributed to the dataset's significant size and intricate underlying patterns.

Evaluation

F1 Score	Decision Tree	k-NN	Logistic Regression
Untuned	0.77	0.75	0.45
Nested CV	0.77	0.79	0.67



The **K-NN** model with **Nested-Cross validation** is the best model that is easy to understand while avoiding overfitting

Deployment of Credit Score Deployment Model

Banks operate in an environment where accurate risk assessment is pivotal for profitability and sustainability. Predictive models, especially related to credit scoring, offer better risk management and decision making processes, which in turn leads to strategic growth and cost savings.

Loan Approvals
Predictive analysis offers insights into a client's repayment ability, streamlining the loan decision process.
Implication: Facilitates quicker, data-driven loan decisions, minimizing potential bad loans.

Credit Card Issuance
The model gauges how a client might manage credit card debt, guiding credit card issuance and limit decisions.
Implication: Optimizes the bank's credit card portfolio, reducing potential defaults and delinquencies.

Interest Rate Personalization
By determining individual credit risk, banks can offer tailored interest rates to clients.
Implication: A competitive edge in attracting creditworthy clients, while balancing risk with higher rates for riskier profiles.

Key Deployment Issues & Ethical Considerations

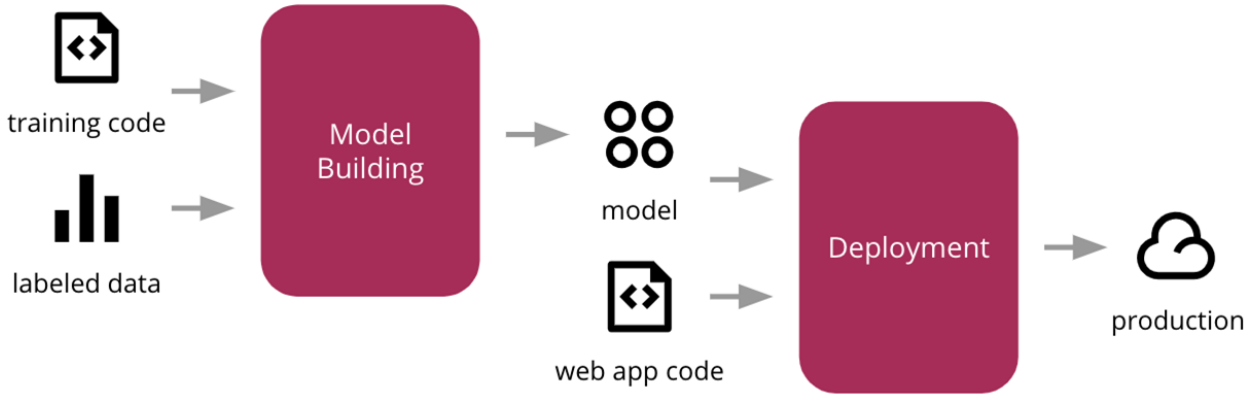
Model Monitoring: Track performance for drifts and changes.

Scalability: Ensure infrastructure meets demand.

Bias & Discrimination: Avoid perpetuating historical biases or targeting protected groups.

Transparency: Make model decisions clear and interpretable.

Data Privacy: Safeguard sensitive information and ensure informed consent.



Risk & Mitigation

- **Unintended Model Biases**
  Mitigation: Utilize fairness-enhancing interventions during model training, and perform regular audits with fairness metrics. Seek diverse input to identify blind spots.
- **Degraded Performance over Time**
  Mitigation: Implement regular model re-training sessions with fresh data, and have monitoring systems in place to alert for significant performance drops.
- **Unauthorized access to personal data**
  Mitigation: encrypt sensitive data both at rest and in transit, coupled with strict access controls that limit data availability to authorized personnel.

THANK YOU