LOAN DEFAULT PREDICTION

A loan default prediction model is a risk management tool that assesses the credit worthiness of a loan applicant by estimating their probability of default based on historical data. It uses numerical tools to rank order cases using data integrated into a single value that attempts to measure loan risk, aiding in informed lending decisions.

01 Overview of the problem

02 Build input data

03 Data processing & analysis

04 Model building

05 Deployment plan

1

Overview of the problem

Introduction of loan default prediction model

1. Problem Modeling

Business Objectives

20%

Loan defaults result in financial losses for banking institutions*

Many financial banks use ML for loan default prediction

trustingsocial

FinScore

kt

Lenddo

It becomes important for banks to understand the behavior of the **customer** before they can take action to lend money to people for different purposes.

* *Theo dữ liệu báo cáo World Bank 2020*

Data Sources

Personal Information: geographical information, address information

Loan Application Data: Applicant's personal information, loan amount requested, purpose of the loan.

Financial History: Credit score, existing debts, payment history, income details.

Banking Information: Account balances, transaction history, overdraft frequency (from open banking APIs).

1. Problem Modeling

Overview of problem requirements

Loan Default Prediction for customers in need of consumer credit loans. Model illustration:

Input

Person_id

Additional information for customer authentication

Full name
Date of Birth
Address

ML Model

Output

The likelihood of loan defaults

Forecasting the probability of customer default 0/1

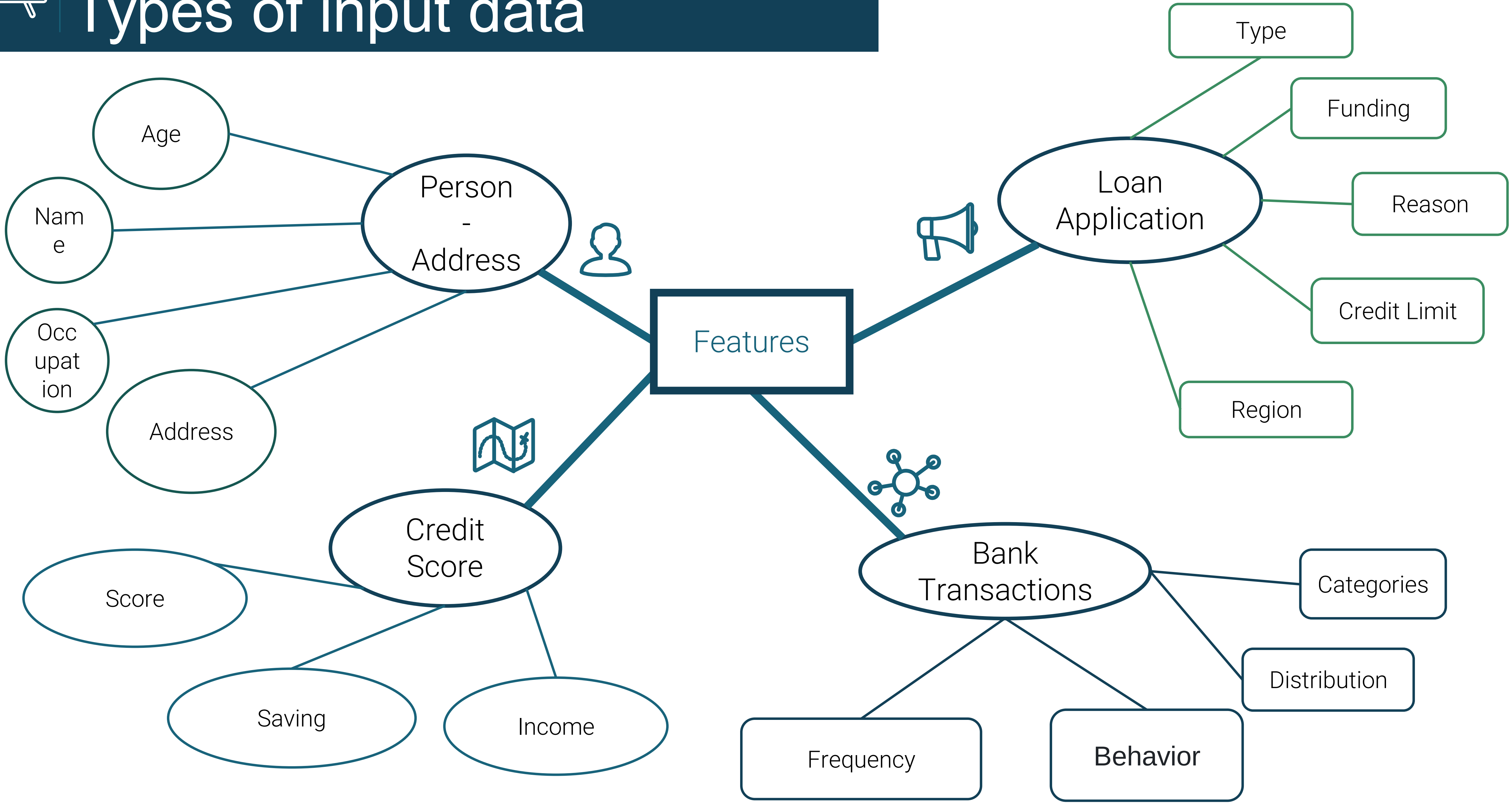
Supporting decision-making process in giving loans

2

Features Engineering

2. Input

Types of input data



Description of input data

The total number of input variables is 121 variables and there are weekly data updates for the model to ensure continuous model improvement. The problem uses 6 different time points:

During the week (Monday - Friday)

Work (9:00 a.m. – 5:00 p.m.)

Rest (5:00 p.m. – 11:00 p.m.)

Night (11:00 p.m. – 7:30 a.m.)

Weekend (Saturday - Sunday)

Work (9:00 a.m. – 5:00 p.m.)

Rest (5:00 p.m. – 11:00 p.m.)

Night (11:00 p.m. – 7:30 a.m.)

- All variables are aggregated according to these 6 milestones to best distinguish customer behavioral characteristics.
- Data were extracted for 2 consecutive months: Increasing the observation time aims to increase the stability of observations, and more clearly see the financial trends of customers.

2. Input

Personal Data

- Age
- Account Age
- Address
- Address Duration
- Region
- Occupation
- ....

An earpiece with complete and clear identification information will often have higher reliability

2. Input

Convert currency

Bank Balance

- Amount
- Varies Currency

Credit Score

- Amount
- Varies Currency

Loan applications

- Total Debt
- Saving
- Monthly Income

Bank Transactions

- Amount
- Balance
- Varies Currency

From the data, we also see that each account might have a different currency balance, so we will convert all amounts to GBP with a real-time rate.
It will ensure consistent in the scale of money between different types of data.

Bank transctional data

From the provided bank transactional data, we can can derive various attributes that may be indicative of a customer's financial behavior and potential default risk.

1.Transaction Amount Statistics:

- Mean transaction amount over the last 7 - 30 - 60 days
- Median transaction amount over the last 7 - 30 - 60 days
- The standard deviation of transaction amount over the last 7 - 30 - 60 days
- Total transaction amount over the last 7 - 30 - 60 days

2.Transaction Frequency:

- Number of transactions over the last 7 - 30 - 60 days
- Number of transactions during different time periods (e.g., business hours, weekends)

2. Input

Bank transactional data

From the provided bank transactional data, we can can derive various attributes that may be indicative of a customer's financial behavior and potential default risk.

3.Transaction Categories:
• Number of different transaction categories (e.g., groceries, utilities, entertainment) - bank_transaction_code
• Frequency of transactions in some main category

4.Transaction Patterns:
• Regularity of transactions (e.g., presence of recurring payments) Time between transactions (e.g., average time between transactions)
• Time since the last transaction

2. Input

Bank transctional data

From the provided bank transactional data, we can can derive various attributes that may be indicative of a customer's financial behavior and potential default risk.

5.Account Balance Trends:
• Average account balance
• Minimum and maximum account balance
• Account balance volatility (e.g., standard deviation of account balance)

6.Distribution of transactions across different locations (e.g., city, country)
• Frequency of transactions in each location

7.Payment Behavior:
• Proportion of successful payments
• Proportion of failed payments
•Proportion of successful Credit payments
• Proportion of failed Debit payments

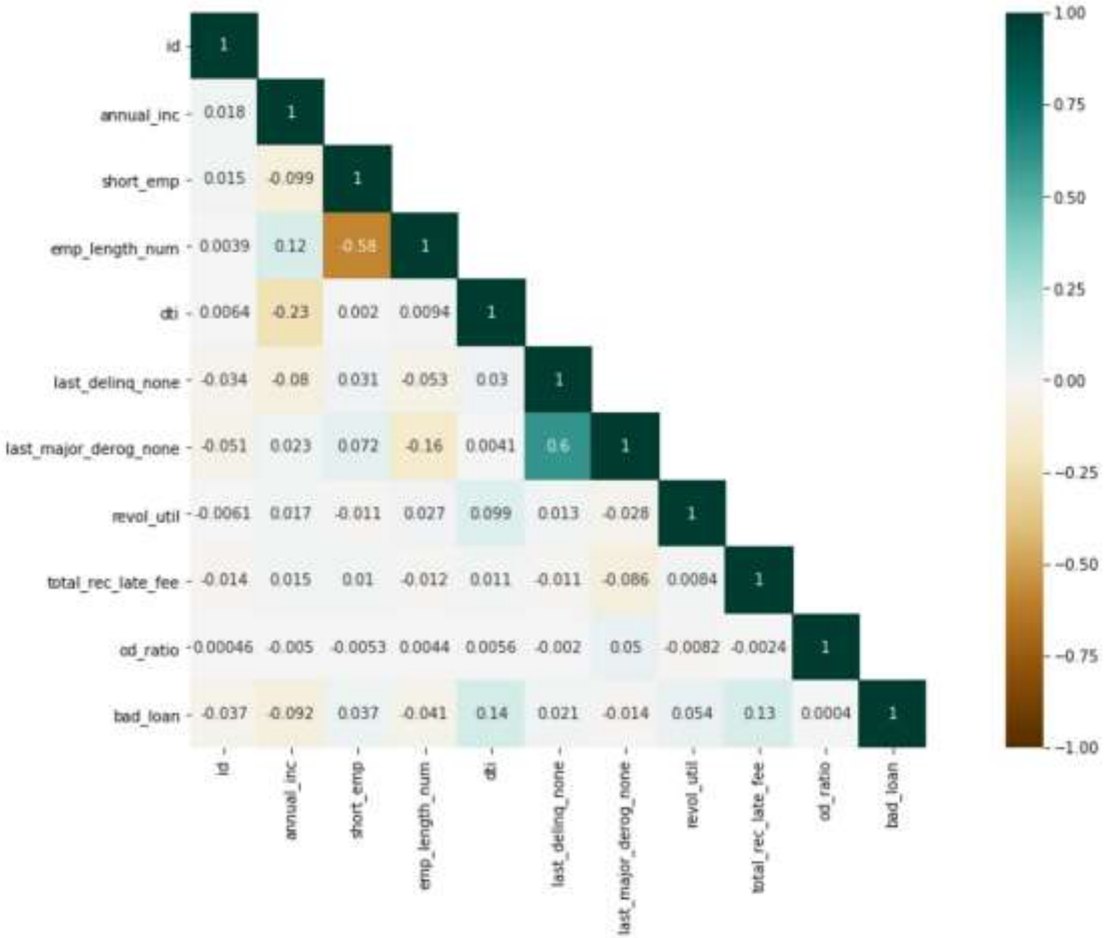
3

Data processing & analysis

Since the provided is synthetically duplicated with the same values which cannot be used to train Machine Learning models for fraud detection.

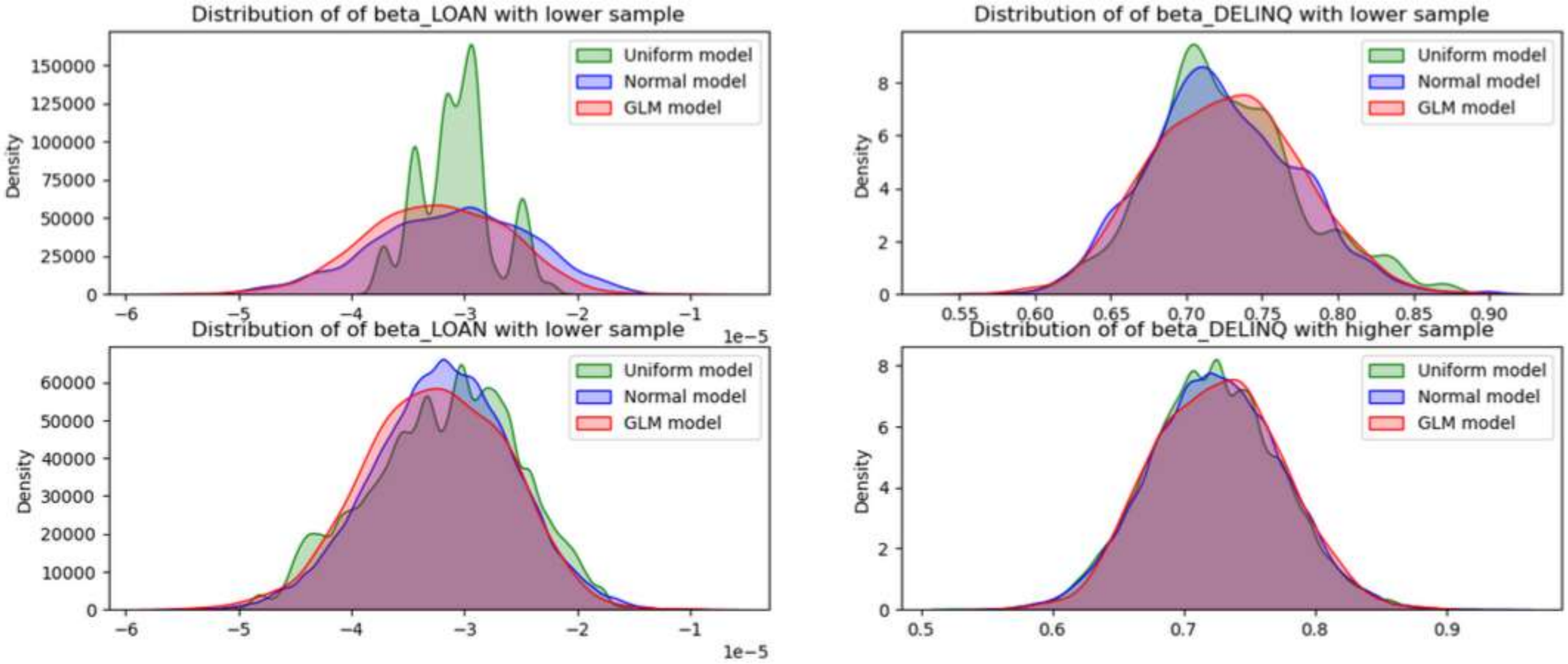
From here, we will suggest some insights and highlight techniques based on the common knowledge and state-of-the-art Machine Learning models for loan default prediction.

Correlation Check



Sample visualization: The top 10 variables best correlated with the bad/good debt label have coefficients ranging from 0.28 to 0.35. It can be seen that these fields have a pretty good linear relationship with the bad/good debt labels.

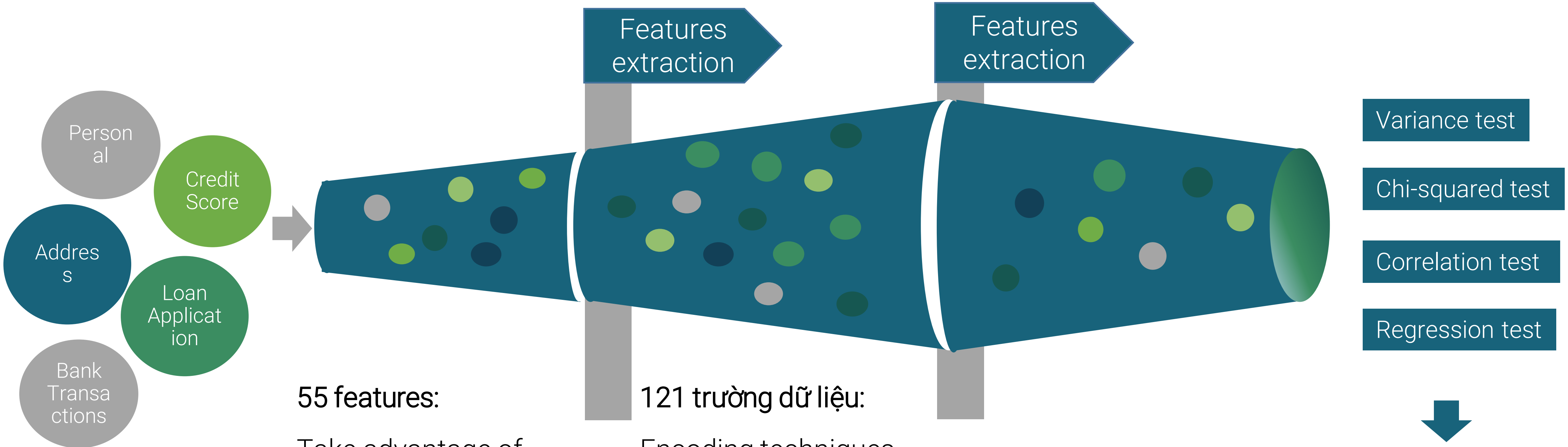
Data Distribution



Example visualization: Visualizing distributions helps in understanding the overall structure of the data and identifying any patterns or anomalies. It provides a visual summary of the data's central tendency and shape, and identifying Outliers:

Data Selection Mindset

The data creation mindset will be in the direction of maximizing exploitable information, including three steps: creating basic data (**raw data**), exploiting information fields from basic data (**Features extraction**), using transformation techniques to create secondary data (**Features Extraction**).



55 features:

Take advantage of information and existing data sources

121 trường dữ liệu:

Encoding techniques (max, min, median, mean, entropy,...)

Data that has no value for scoring will be eliminated during the processing and cleaning process, or will be automatically screened and selected by the model.

Feature Engineering

- Using data for 3 months can increase the prediction results. The author recommends using 6 consecutive months of data to increase stability.

- With numeric data: Group into groups according to different months, turn into data that is the mean and std of these data. These actions increased the prediction result (ginin index) by about 0.015 compared to using the original data.

- For categorical data: Encoder using a regular label encoder. According to the author's experience, encoding using other methods such as onehotencoder, target encoder does not improve accuracy if using boosting algorithm.

4

Model Building

Data modelling

Overview of model flow

Pipeline Storage System

Prepare Data → Train Model → Deploy Model → Monitor Model

Features Extraction

Data Storage

Data Preparation
- Explore data
- Missing cleaning
- Outlier cleaning
- Label Encoding
- Normalization

Model Training & Validation
- Model Selection
- Hyper-parameter Tuning
- Model testing

Model Deployment
- Intergrate into system
- Batch Scoring

Inference
- Tracking performance
- New real-life data

- Set rules to knock-out records
- Report data processing step by step

- Usually store versions of model
- Daily track performance of model

- Store **history of customer scoring**
- Track data updates in database

4. Model Building

Supervise model operation

Scoring data will be updated once a day to enrich the data. The richer the data gets, the better the model performs

Monitor input data

Check for null value

Check for abnormal values

Monitor the processing process

Check variable values (features)

Check distribution and abnormality of variables

Database

Put in ML Model

Monitordistribution RISK_FLAG

Check for abnormal fluctuations in the distribution

Point encryption (users cannot arbitrarily adjust, edit, or delete)

Push data to Beta version

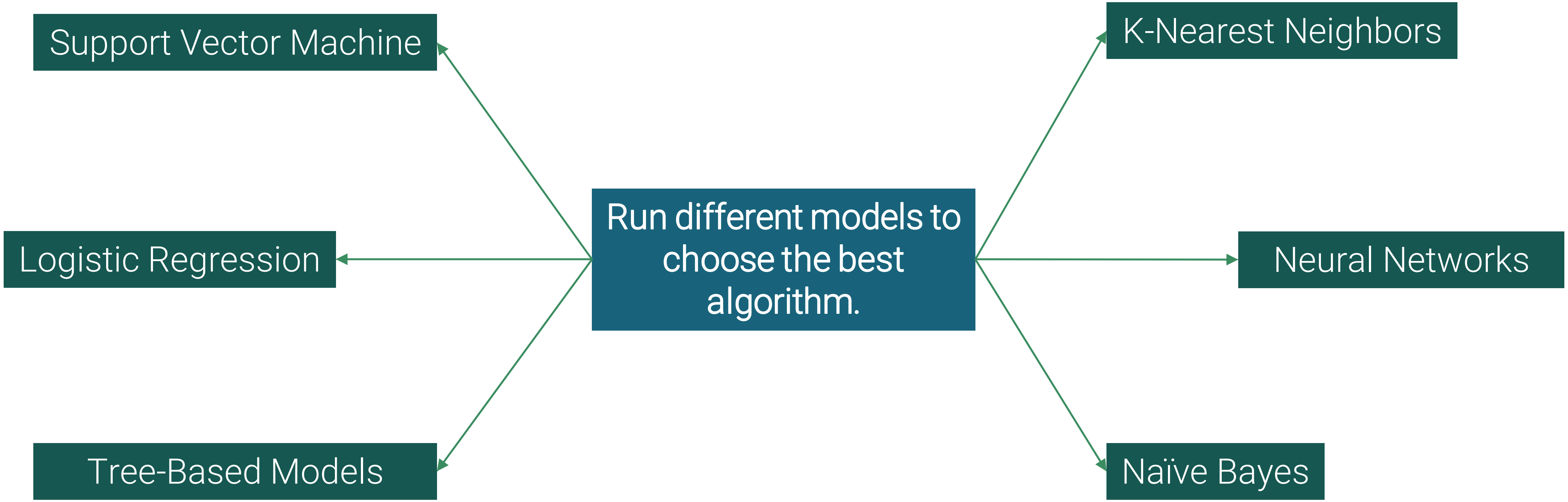
Check the score data again before posting to the official version

Save history results and update as model input data

Official version

The model is continuously monitored and updated helps learn new trends such as market fluctuations, epidemics, etc., thereby adjusting to match the trends.

Model Selection



Simple algorithms such as Logistic Regression or Naïve Bayes are used as baseline.
For each algorithm, run Baysian research CV or Grid Search to find the set of parameters that best fits that model

Model Selection

Support Vector Machine

Logistic Regression

Neural Networks

Tree-Based Models

- LightGBM
- CatBoost
- XGBoost
- Decision Tree
- Random Forest

Based on model performance in practice

Boosting algorithms in the family of Tree-Based models such as LightGBM, XGBoost, CatBoost always achieve the best results with classification problems.*

Based on the strengths of each model and expert experience

Random Forest/ XGBoost are chosen to build the model

** According to the papers referred to in the appendix*

Basically the Random Forest, LightGBM, XGBoost, CatBoost algorithms are the same, belonging to the same family of boosting algorithms: Using many weak Tree Decisions to create a Robust Decision and limiting overfitting.

Model Strengths

Strengths of Random Forest/XGBoost

Consumes few resources, easily runs on Hadoop Spark (server) or personal computer (local).

Good results on a variety of datasets with very few parameter settings.

Good limitation of overfitting problem.

Easily handle missing data, non-numeric data, and skewed label ratios.

Advantages of K-Folds*

Applying K-Folds alongside LightGBM aims to provide the most objective and stable results

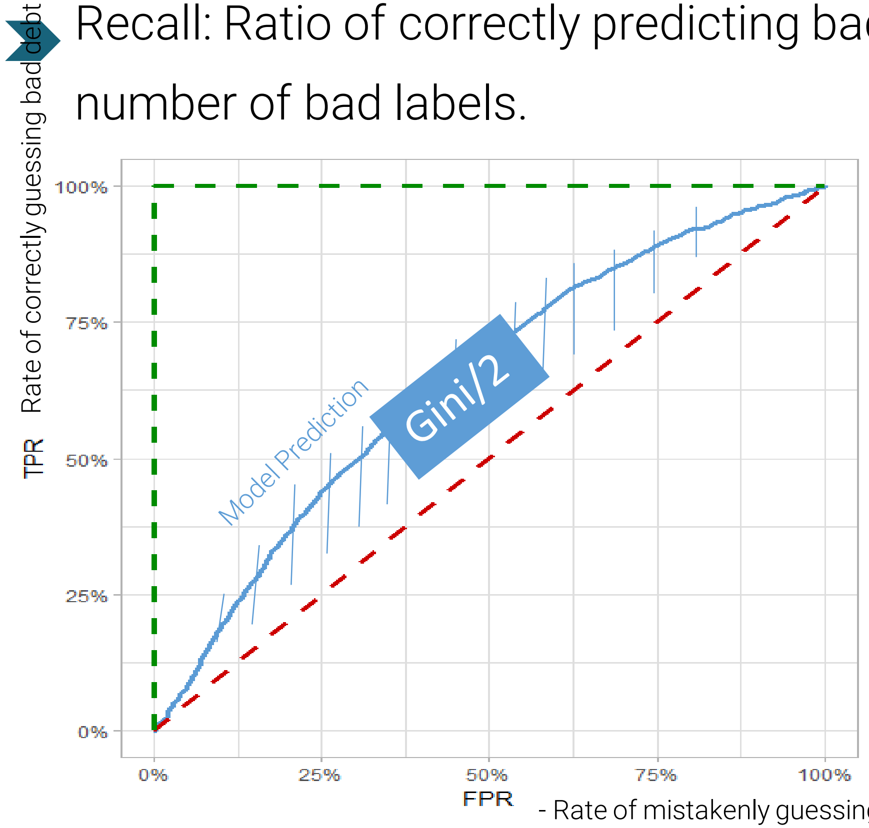
Minimize interference factors caused by the process of cutting the training and test sets.

Method: Randomly divide the data set into 5 equal parts. In turn train on 4 parts and test the results on the remaining set

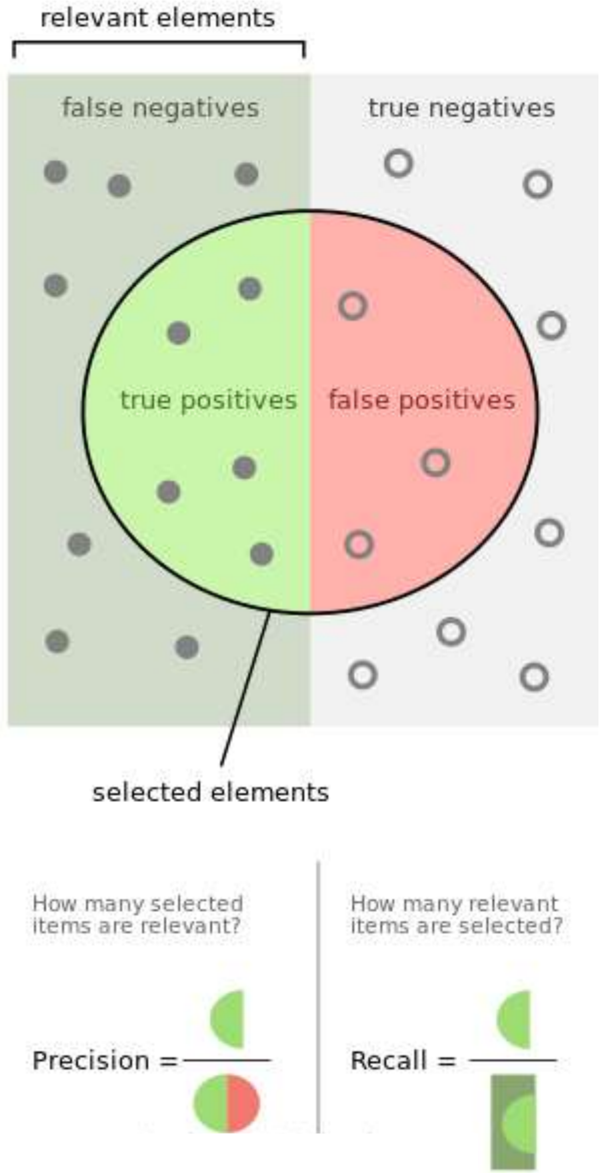
** Chi tiết kết quả chạy K-fold xem phần phụ lục*

Evaluation Metrics

- Gini: This value ranges from -1 to 1, corresponding to completely wrong guessing and absolutely correct guessing.
- Recall: Ratio of correctly predicting bad labels out of the total number of bad labels.

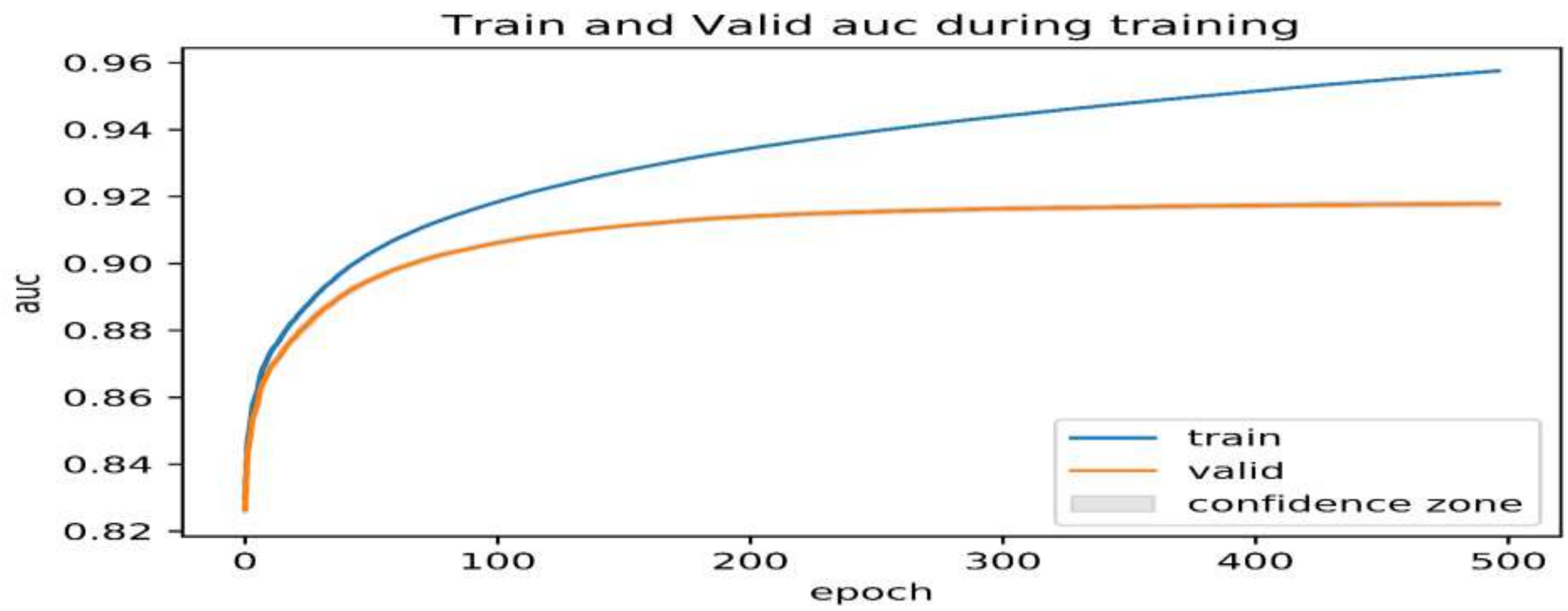


- Rate of mistakenly guessing good loan as bad loan



https://en.wikipedia.org/wiki/Precision_and_recall

Model Evaluations



When the dataset is large enough, the ensemble tree-based models performed on different training and test sets all produce results with very small standard deviations, proving the very high stability of these models.

Model Prediction

- It is possible to change the structure of the confusion matrix by changing the model threshold on good/bad loan to better suit each situation:
- Minimize loan debt: In case we do not have much capital, or have a low risk appetite, the high threshold will be reduced, but the limitation is that having a lot of good loan application will not be considered for loans due to strict standards.
- Increase disbursement rate: In case we have abundant capital and high risk appetite, consider a low threshold but the limitation is that the bad loan ratio is at risk of increasing.

5

Model Deployment

The deployment plan is presented in the Deployment Strategy.pdf file

Deliverables

- Technical Report.pdf - A high-level and concise report on the approach, architecture decisions, development process, and model insights.
- README.md - A comprehensive guideline and analysis report detailing the approach, architecture decisions, development process, and model insights.
- Loan_Default_Prediction.ipynb - A predictive model for identifying high-risk loan applicants.

Deployment Strategy.pdf A deployment strategy with monitoring and updating protocols.

References

- *Lane, M., Carpenter, L., Whitted, T., Blinn, J.: Scan line methods for displaying parametrically defined surfaces. Communications ACM 23(1) (1980)*
- *Ding, C.H.Q., Peng, H.: Minimum redundancy dữ liệu selection from microarray gene expression data. J. Bioinformatics and Comp. Biol. 3(2) (2005) 185–206*
- *Prediction of Socioeconomic Levels Using Cell Phone Records*
- *Credit Scoring for Good Enhancing Financial Inclusion with Smart(2019)*
- *Business Intelligence Applications and Data Mining Methods in Telecommunications: A Literature Review https://core.ac.uk/tải xuống/pdf/71542208.pdf*
- *And many other at: http://confluence.digital.vn/x/ZyRDAg*