Credit Card Approval Prediction

DS105 – Final Project Presentation

By Jia Yang

Table of Content

Problem Statement & Goal

Problem Statement

Banks heavily rely on credit score to assess applicant creditworthiness that may **lose it predictive power** due to large economic fluctuation

Credit score's creditworthiness **do not paint a complete picture** of the applicant such as their personal information. Only rely on historical data such as payment history and credit utilization

Goal

To **build a Machine Learning Model** to predict "good" or "bad" credit card applicant **based on the collected personal information's** from the applicant with will not lose it predictive power.

Dataset Review

Dataset Review

Datasets

1. Application_record.csv
2. Credit_record.csv

https://www.kaggle.com/rikdifos/credit-card-approval-prediction

Description

- Datasets are **connected by customer IDs**.
- **Application_record.csv** contains applicant personal information, can be use as **features**. *(Total 18 columns and 439k rows)*
- **Credit_record.csv** records the applicant behaviours of credit card, can be use as **label**. *(Total 3 columns and 1.05m rows)*

Dataset Review

Variable Types - Application_record.csv

No.	Feature name	Description	Variable Type	Data Type	Variable Category
1	ID	Client number	-	Numeric	Continuous
2	CODE_GENDER	Gender	Predictor	Numeric	Categorical
3	FLAG_OWN_CAR	Is there a car	Predictor	Character	Categorical
4	FLAG_OWN_REALTY	Is there a property	Predictor	Character	Categorical
5	CNT_CHILDREN	Number of children	Predictor	Numeric	Continuous
6	AMT_INCOME_TOTAL	Annual income	Predictor	Numeric	Continuous
7	NAME_INCOME_TYPE	Income category	Predictor	Character	Categorical
8	NAME_EDUCATION_TYPE	Education level	Predictor	Character	Categorical
9	NAME_FAMILY_STATUS	Marital status	Predictor	Character	Categorical
10	NAME_HOUSING_TYPE	Way of living	Predictor	Character	Categorical
11	DAYS_BIRTH	Birthday count backwards from current day (0), -1 means yesterday	Predictor	Numeric	Continuous
12	DAYS_EMPLOYED	Start date of employment count backwards from current day(0). If positive, it means the person currently unemployed.	Predictor	Numeric	Continuous
13	FLAG_MOBIL	Is there a mobile phone	Predictor	Numeric	Categorical
14	FLAG_WORK_PHONE	Is there a work phone	Predictor	Numeric	Categorical
15	FLAG_PHONE	Is there a phone	Predictor	Numeric	Categorical
16	FLAG_EMAIL	Is there an email	Predictor	Numeric	Categorical
17	OCCUPATION_TYPE	Occupation	Predictor	Character	Categorical
18	CNT_FAM_MEMBERS	Family size	Predictor	Numeric	Continuous

Both datasets will be connected with Client Number

Dataset Review

Variable Types - Credit_record.csv

No.	Feature name	Description	Variable Type	Data Type	Variable Category
1	ID	Client number	-	Numeric	Continuous
2	MONTHS_BALANCE	The month of the extracted data is the starting point, backwards, 0 is the current month, -1 is the previous month, and so on	Predictor	Numeric	Categorical
3	STATUS	Payment Status 0 : 1-29 days past due 1 : 30-59 days past due 2 : 60-89 days overdue 3 : 90-119 days overdue 4 : 120-149 days overdue 5 : Overdue or bad debts, write-offs for more than 150 days C : paid off that month X : No loan for the month	Target	Character	Categorical

Both datasets will be connected with Client Number

Label for credit card approval based on the acceptable past due range

Dataset Review

Snapshots - Appication_record.csv

	ID	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	NAME_INCOME_TYPE	NAME_EDUCATION_TYPE
0	5008804	M	Y	Y	0	427500.0	Working	Higher education
1	5008805	M	Y	Y	0	427500.0	Working	Higher education
2	5008806	M	Y	Y	0	112500.0	Working	Secondary / secondary special
3	5008808	F	N	Y	0	270000.0	Commercial associate	Secondary / secondary special
4	5008809	F	N	Y	0	270000.0	Commercial associate	Secondary / secondary special

NAME_FAMILY_STATUS	NAME_HOUSING_TYPE	DAYS_BIRTH	DAYS_EMPLOYED	FLAG_MOBIL	FLAG_WORK_PHONE	FLAG_PHONE	FLAG_EMAIL
Civil marriage	Rented apartment	-12005	-4542	1	1	0	0
Civil marriage	Rented apartment	-12005	-4542	1	1	0	0
Married	House / apartment	-21474	-1134	1	0	0	0
Single / not married	House / apartment	-19110	-3051	1	0	1	1
Single / not married	House / apartment	-19110	-3051	1	0	1	1

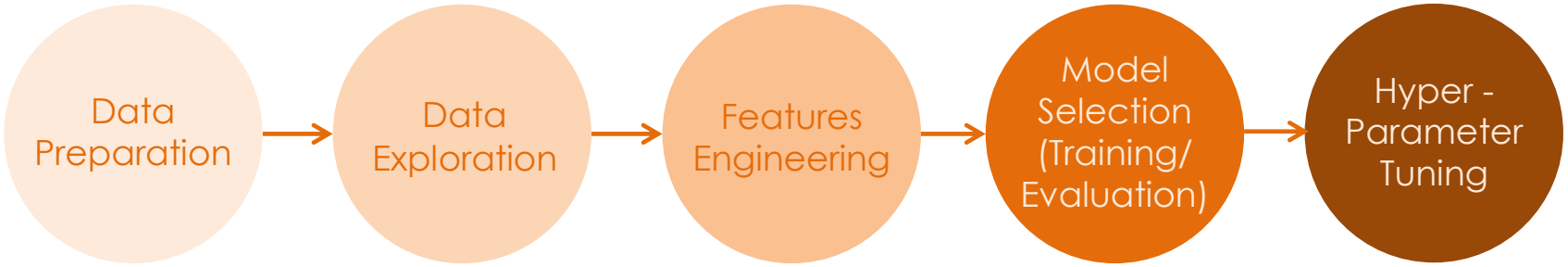
Dataset Review

Snapshots - Creidt_record.csv

	ID	MONTHS_BALANCE	STATUS
0	5001711	0	X
1	5001711	-1	0
2	5001711	-2	0
3	5001711	-3	0
4	5001712	0	C

Machine Learning Approach & Challenges Anticipated

Machine Learning Approach & Challenges Anticipated



Challenges

- Null values treatment
- Duplicate records
- Joining of datasets

- Relationship between variables

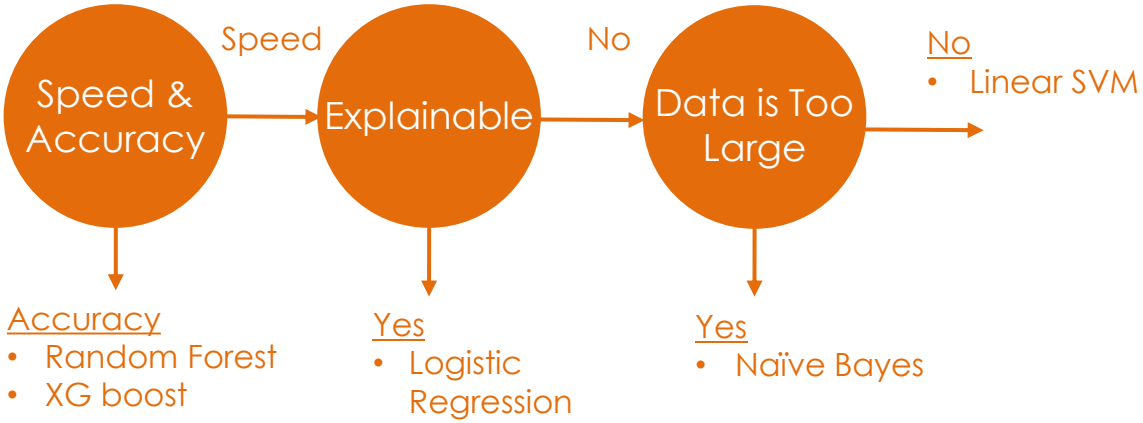
- Create Label
- Create new meaningful features
- Reduce unused features
- Data unbalance.
- Encode categorical features
- Scale overall dataset

- Choose the best Classification ML model based on different evaluation method

- Choose the best hyper parameter

Machine Learning Approach & Challenges Anticipated

Classification Models Selection



Evaluation Methods

- Confusion Matrix
  - Accuracy
  - Recall
  - Precision
  - F1-Score
  - ROC / AUC


Sub-Goals

Sub-Goals

1. Method of creating label from credit_record.csv?
2. Encoder used for all categorical features?
3. Clear segregation between the “good” and “bad” credit card applicants?
4. The best classification machine learning model for prediction?
5. The best hyper-parameters to give the best prediction result?
6. The feature that give the major contribution to the prediction?


Thank You
https://github.com/jiayang243/CreditCardApproval