Default Risk

Peining Fan | Yidan Gao | Dimo Liao | Shunshun Miao | Chuchen Xiong | Yuting Xin

Situation

- Business Vision: Home credits offers affordable financial loans and has over 100M customers and clients located throughout 9 countries.
- Our Objective: Assist Home Credit to expand customer base on the underestimated populations.

Complication

Situation: Previously, Home Credit relied on internal credit history of applicants. However, the new entrants with limited credit history are treated highly biased and excluded.

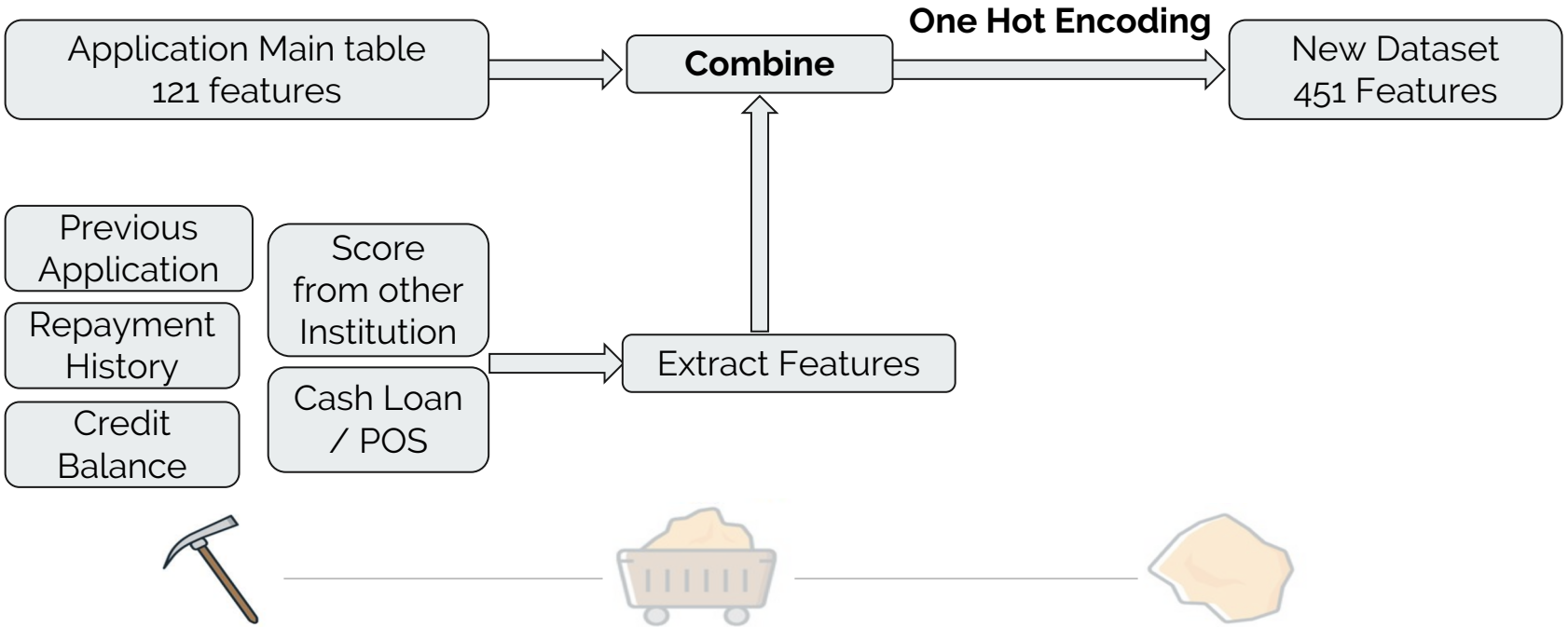
Resolution: Applied external demographic and historic data to broaden financial inclusion for the credit-limited population.

Approach

- Feature Engineering
- Data Cleaning
- Feature Selection

Approach - Feature Engineering

- Task: Extract features from five supplementary files and merge into main files



Approach - Data Cleaning

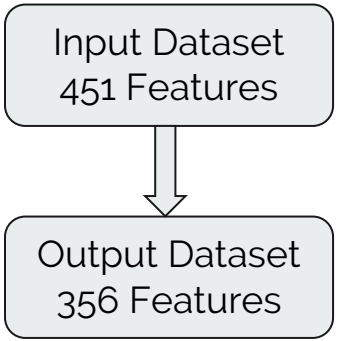
- Missing Values Processing
- Outliers and Anomalies Modification

Approach - Data Cleaning

- Missing Values Processing

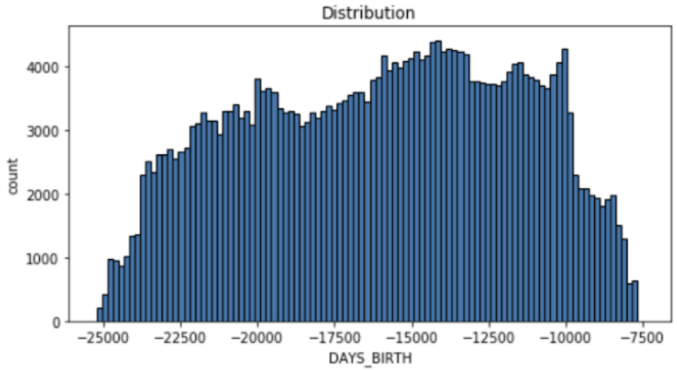
➔ Remove the columns with 30% NAs or more

➔ Fill the NA with average value



Approach - Data Cleaning

- Outliers and Anomalies Modification

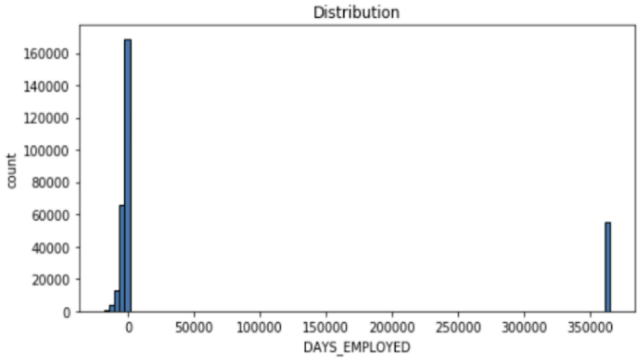


“Good” Distribution

No Apparent Outliers

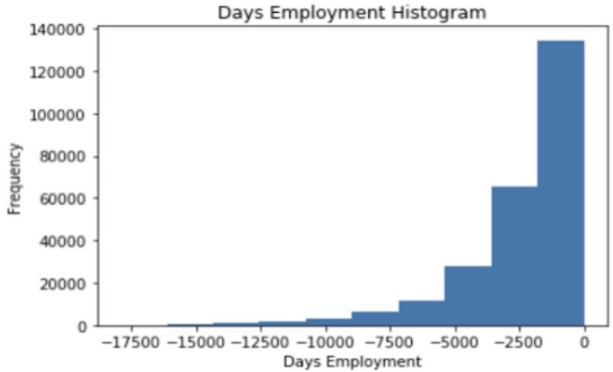
Approach - Data Cleaning

- Outliers and Anomalies Modification



➔ Replace outliers/anomalies with NA value

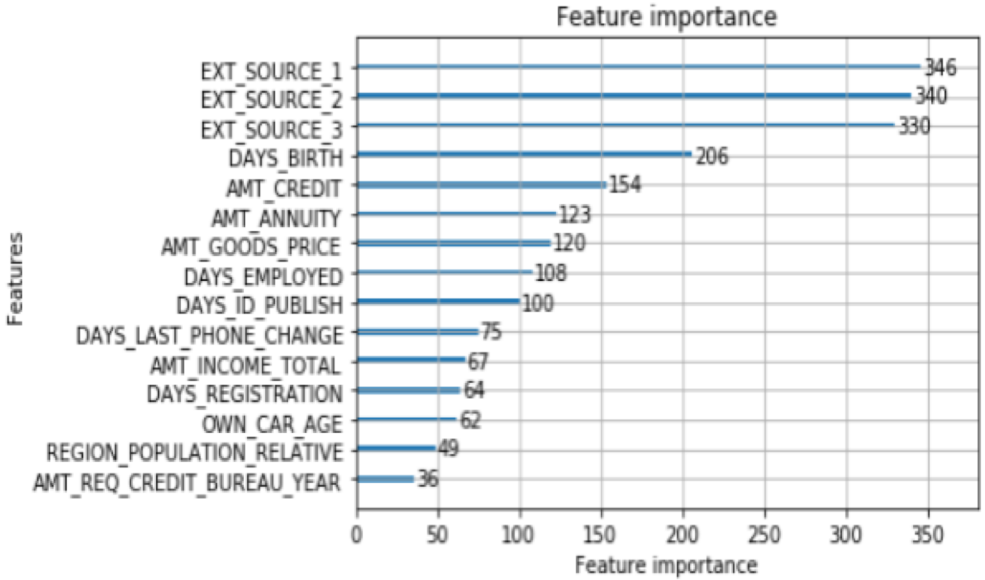
➔ Fill “new” NA with mean value



“Bad” Distribution

“Good” Distribution

Approach - Feature Selection



356 Features

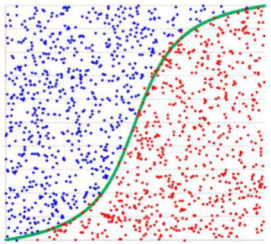
Correlation → 29 and 100 Features

LGBM → 30 and 98 Features

PCA → 12 Features

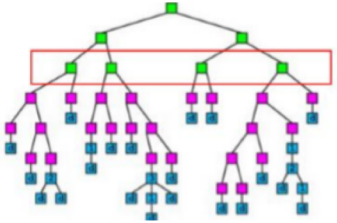
Model Comparison

Logistic Regression



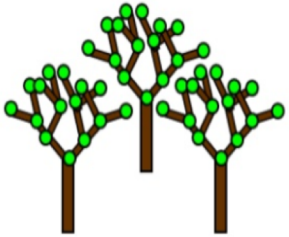
- Binary Classification
- Sigmoid Curve

XGBoost



- Parallelized in information gain
- Less train time

Random Forest

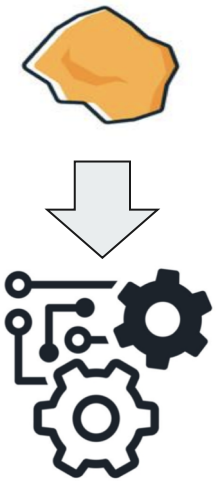


- Parallel ensemble
- Feature importance is weighted more

LightGBM

- Decision Tree based
- Leafwise growth
- Handle categorical features

Predication and Preparation



Performance Score		Logistics Regression	Random Forest	Light GBM	XGBoost
Main table (238 features)		0.59816	0.68146	0.72730	0.71758
Full table (356 features)		0.72434	0.58546	**0.74449**	0.73437
Correlation select features	(100 features)	0.68985	0.58408	0.74329	0.71675
	(29 features)	0.70746	0.70525	0.72399	0.72464
LGBM select features	(98 features)	0.68980	0.58307	0.74287	0.72097
	(30 features)	0.68141	0.5120	0.72537	0.71921
PCA	(12 features)	0.50000	0.50320	0.50416	0.49386
Average Performance		**0.65586**	**0.593503**	**0.70173**	**0.689626**

Score Metric: Area under ROC (It tells how much model is capable of distinguishing between classes)

Impact & Value

Leaking revenue from new applicants are retained.

The unqualified applicants are identified more precisely.

Conclusion & Next Step

Technical Finding

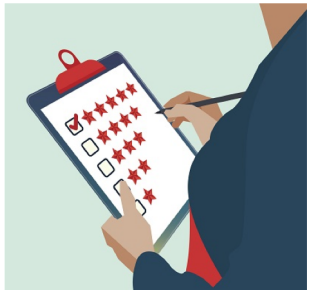
- Only a small number of features are informative
- Light GBM shows best performance

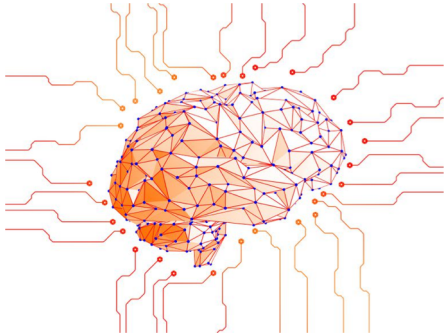
Recommendation

- Understand what external source 1,2,3 means in this context
- Weight more on applicants' age and employment tenure
- Annuity and previous credit history are decisive

Next Step

- Create more domain knowledges features
- Try on complexed model such as Neural Network





Q&A