DSA5101 Project

Bank Marketing Analysis and Prediction

JIANG HAOCHEN A0275715X

LIU HUANSHUO A0275630E

WU SHUANG A0275610J

ZHANG XIAOTING A0275697E

ZHOU YU A0275686J

2023/10/1

National University of Singapore

Table Of Contents

Introduction and Problem Statement

- **Topic:** Bank Customers Deposit Prediction
- **Objective:** Predict whether a bank customer will choose to make a deposit or not.
- **Dataset:**
  - Focuses on bank customers and their deposit behaviors.
  - # of Feature Fields: 16
  - Target Variable (y): Boolean (Deposit or Not)
- **Task Nature:**
  - Binary Classification Prediction
  - Categorical Prediction based on customer dimensions and behaviors.
- **Similarity:**
  - Comparable to Click Through Rate (CTR) Prediction task.

- **Models Used:**
  - DeepFM
  - Random Forest
  - **Model Ensembling**
  - Other relevant algorithms...
- **Exploratory Analysis & Preprocessing**
  - Overview and preparation of the dataset
  - Identification of patterns, anomalies, and trends
- **Model Selection & Performance**
  - Analysis of different models post feature engineering
  - Selection of the ensembling model based on performance metrics
- **Future Directions & Conclusion**
  - Possible improvements and methods for implementation
  - Summary of key findings and takeaways

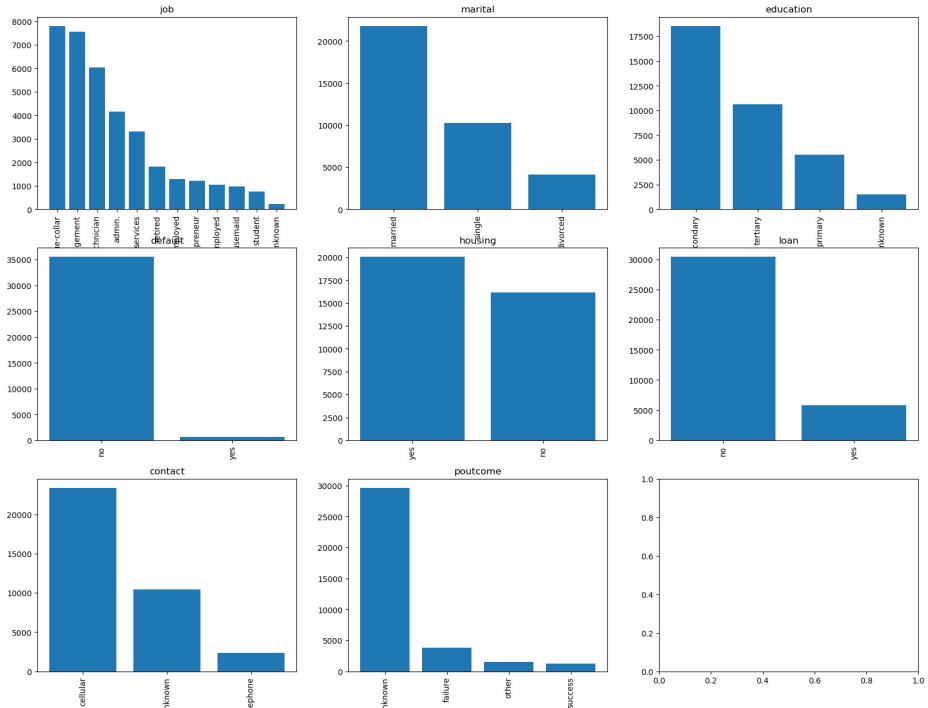
- **Dataset Imbalance:** 31937 : 4231 ≈ 7.5 : 1
- **Model Evaluation Metrics Considered:**
  - Precision, Recall
  - F1 Score, Macro F, Micro F
  - AUC (Area Under the Curve)
  - Logloss
  - Accuracy
- **Selected Primary Metric:**
  - **AUC (Area Under the Curve)**
    - Chosen due to its reliability in the presence of class imbalance.
    - Avoids biases introduced by converting model prediction probabilities into classification thresholds.
    - Maximizing AUC ensures the positive sample score is as high as possible, aligning with the objective of accurately identifying users likely to make a deposit.

Dataset Pre-processing

Numerical attributes

	age	balance	day	duration	campaign	pdays	previous
mean	40.94	1365.06	15.78	258.36	2.77	39.75	0.58
std	10.62	3098.19	8.31	257.19	3.12	99.55	2.38
min	18	-8019	1	0	1	-1	0
max	95	102127	31	4918	63	871	275

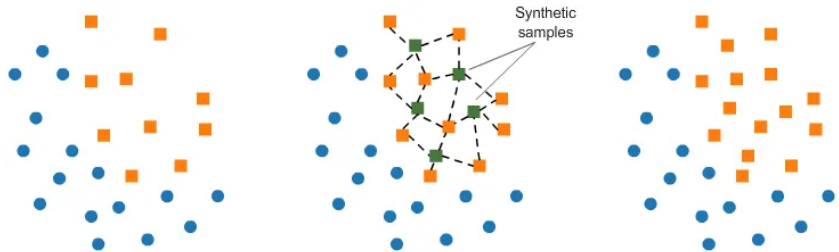
Non-numerical(categorical) attributes



Non-numerical(categorical) attributes

- **Discretization of Continuous Data:**
  - Some algorithms are more suitable for discrete data.
  - Continuous data need to be converted to discrete to use these algorithms.
  - Benefits:
    - Overcomes hidden defects like outliers.
    - Makes model results more robust and stable.

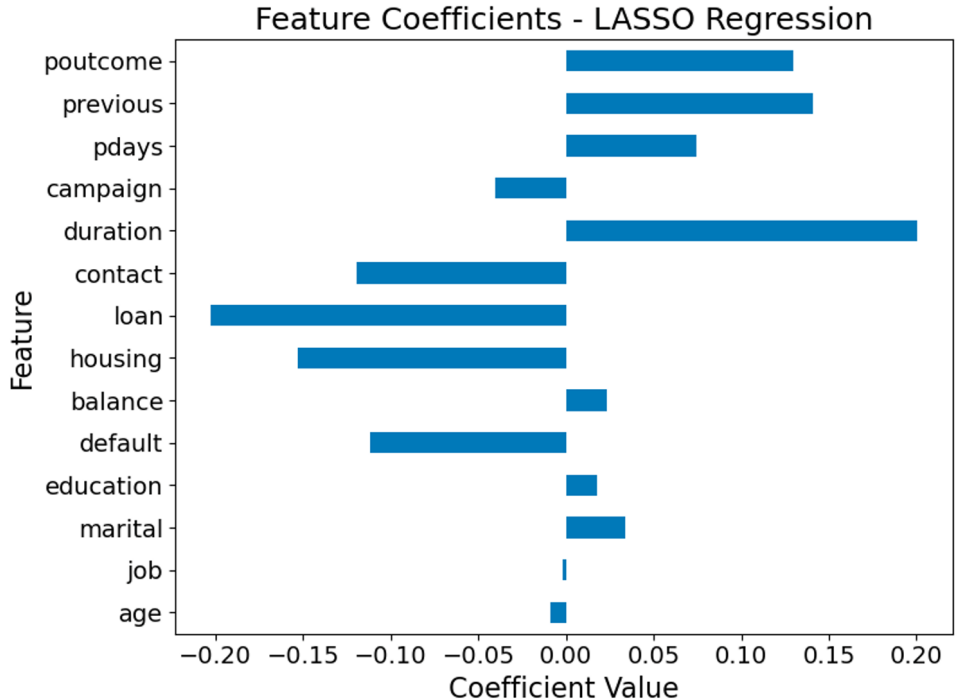
- **Addressing Class Imbalance:**
  - **Oversampling:**
    - Used to address class imbalance issues.
    - Enables better predictions on underrepresented instances.
    - Prevents model bias towards the majority class.
    - Crucial when the minority class contains critical instances.
- **Types of Oversampling:**
  - **Random Oversampling**
  - **SMOTE_NC:**
    - Generates synthetic samples by interpolating feature values between minority class instances.



The principle of SMOTE[1]

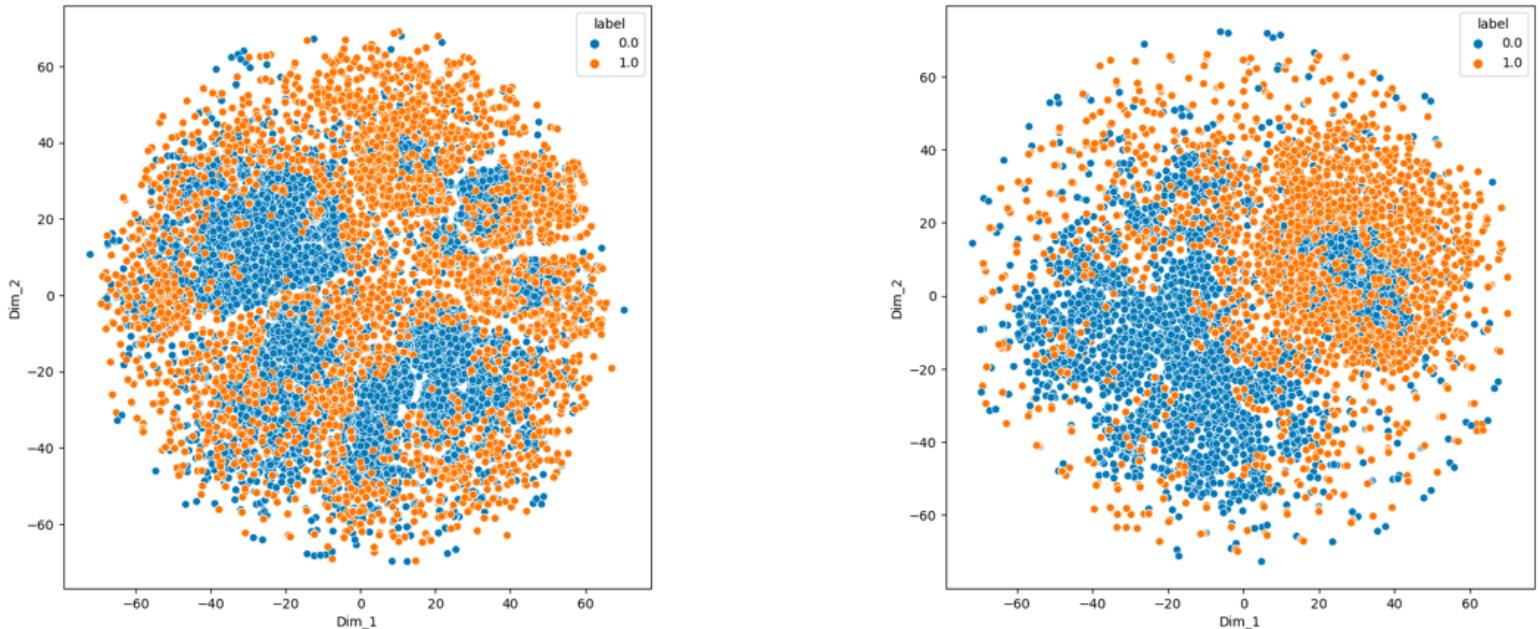
[1]Adapted from Oversampling to remove class imbalance using smote

- **Logistic Regression with l_1 penalty**



Coefficients of Lasso Regression(Logistic Regression with l_1 penalty)

- **T-SNE Visualization**



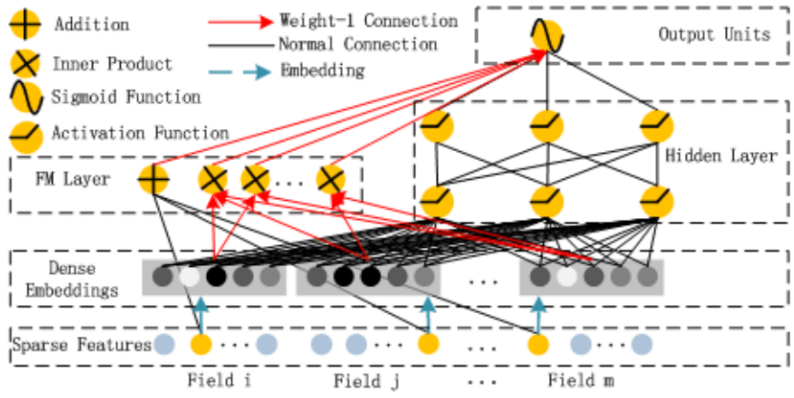
Comparison of the T-SNE visiualization results between data without model selection(left) and data with model selection(right)

Experimental Study and Analysis

- **Models Used:**
  - Logistic Regression
  - Decision Tree
  - Random Forest
  - **DeepFM**
  - XGBoost
  - LightGBM
  - CatBoost
  - **Model Ensembling**

DeepFM

- **DeepFM Overview:**
  - Combination of Wide&Deep and Factorization Machine (FM).
  - Replaces Linear Regression (LR) in Wide&Deep with FM.
- **Advantages over Linear Regression:**
  - FM considers the intersection between features.
  - Improves the model's ability to extract information on the wide side.
- **Deep Model Characteristics:**
  - Can adaptively learn high-order interactions between features.
  - Possesses strong feature extraction and expression capabilities.



Architecture of DeepFM[1]

[1]Adapted from DeepFM: A Factorization-Machine based Neural Network for CTR Prediction

Overall Performance

Model	AUC	LogLoss	F1	micro F	macro F	Precision	Recall
LR	0.8692	0.2746	0.3200	0.8895	0.6295	0.5700	0.2200
DT	0.8713	0.2764	0.5100	0.8980	0.7253	0.5800	0.4500
RF	0.9101	0.2359	0.2400	0.8951	0.5925	**0.7800**	0.1400
DeepFM	0.9310	0.2031	0.6259	0.8937	0.7820	0.5300	**0.7600**
xgb	**0.9379**	0.2019	**0.6302**	0.8983	**0.7856**	0.5500	0.7400
lgb	0.9338	0.2031	0.6199	**0.8999**	0.7811	0.5600	0.7000
catboost	0.9376	**0.1949**	0.6275	0.8942	0.7829	0.5300	**0.7600**

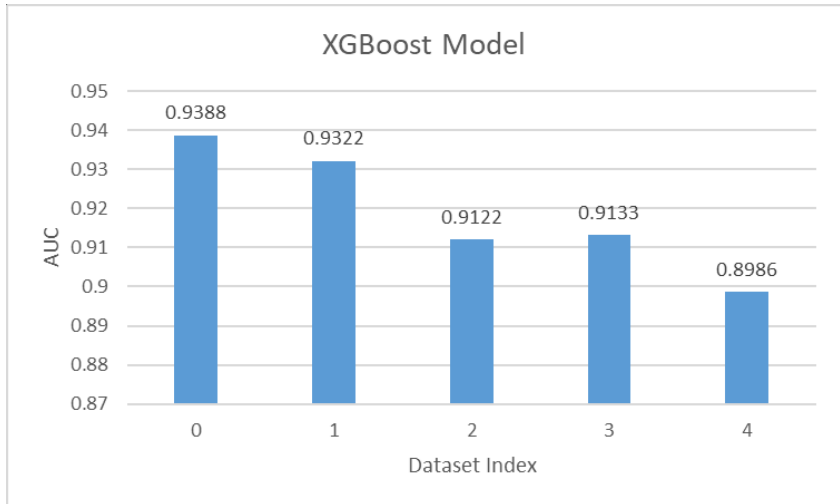
Observations

- **Observation 1: Complex Models vs Simple Models**
  - GBDT-based models and DeepFM significantly outperform simpler models.
- **Observation 2: Best-Performing Model**
  - XGBoost (XGB) is the standout performer, especially with smaller datasets.
  - **DeepFM and LightGBM**
    - Better suited for large-scaled datasets.
  - **CatBoost Performance**
    - Slightly inferior due to the numerous continuous features.
- **Observation 3: GBDT-based Models vs CTR Models**
  - **GBDT-based Models**
    - Perform better than CTR models in this dataset.
  - **CTR Models Limitations**
    - Rely on absent ID-based information, limiting fitting ability.
    - Small dataset size hinders the full potential of deep learning techniques.

Comparison of model performance using different preprocessing methods

Dataset	Caption	age/job/day/month removed	Oversampling	Logist + Lasso
0	no smotenc	no	no	no
1	smotenc	no	smotenc	no
2	smotenc and discretization	no	smotenc	no
3	continous features discretization	no	no	no
4	discretization + smotenc + lasso	yes	smotenc	yes

- **Key Observation:**
  - Best performance consistently achieved on dataset 0 (no preprocessing).
- **Possible Reason:**
  - Inherent cleanliness of dataset 0; excessive processing introduced noise.
- **Conclusion & Implication:**
  - Minimal preprocessing is optimal for clean datasets to maintain data integrity and avoid performance deterioration.



AUC of XGBoost on Dataset 0-4[1]

Project Accomplishments Overview

- **Cost-Benefit Analysis:**
  - Low cost for reaching a customer.
  - Significant benefits from customer deposits.
- **Strategic Priority:**
  - Higher recall value by lowering the model threshold.
- **Model Ensembling:**
  - Employing model ensembling to combine various models.
- **Outcome:**
  - Improved results as depicted in the table.

- Our final model is **optimal** on almost all metrics
- Achive **balance** between precision and recall

Model	AUC	LogLoss	F1	micro F	macro F	Precision	Recall
DeepFM	0.9310	0.2031	0.6259	0.8937	0.7820	0.5300	**0.7600**
xgb	0.9379	0.2019	0.6302	0.8983	0.7856	0.5500	0.7400
lgb	0.9338	0.2031	0.6199	**0.8999**	0.7811	**0.5600**	0.7000
catboost	0.9376	0.1949	0.6275	0.8942	0.7829	0.5300	**0.7600**
Ensembling	**0.9404**	**0.1946**	**0.6374**	0.8985	**0.7892**	0.5500	**0.7600**

Future Enhancement Possibilities

- **Further Analyses:**
  - Conduct analyses related to feature importance.
- **Model Introduction:**
  - Consider employing RIM[1], a state-of-the-art model in Recommender Systems CTR Prediction task.

[1] Retrieval & Interaction Machine for Tabular Data Prediction

Conlusion

- **Summary of Exploration:**
  - Insightful findings in Bank Marketing Analysis and Prediction through the DSA5101 Project.
  - Successful implementation and comparison of various models.
  - Achieved optimal balance in precision and recall with ensembled model.
- **Key Insights:**
  - Superiority of complex models like GBDT-based models and DeepFM, especially XGBoost in smaller datasets.
  - Importance of minimal preprocessing in maintaining data integrity.
- **Strategic Impact:**
  - Emphasis on substantial benefits from customer deposits and prioritizing higher recall values.
  - Aid to financial institutions in refining strategies and informed decision-making.