COMP9417

Presentation

UNSW
SYDNEY

Group 42

Roger Li (z5470193) (Presenter)

Weihao Zhou (z5463147)

Xuda Chen (z5527738)

Hao Zhu (z5503328)

Introduction & Experiment Environment

Problem Part 1: 28-class classification task

- Exploratory data analysis
- Model & method selection
- Row sampling + feature selection + Bayesian hyperparameter tunning
- Ensemble strategy & final prediction

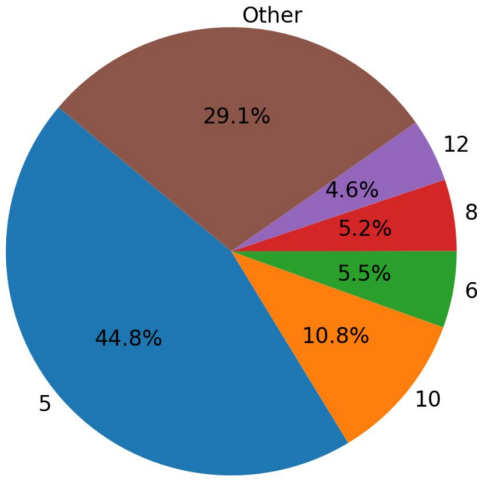
Problem Part 2: Distribution shift

- Data analysis based on definition
- Corresponding strategies

Device	GPU T4 x 2
Execution platform	Kaggle
Random Seed	42
Main Metric	Weighted Cross-Entropy Loss
scikit-learn Version	1.2.2
imbalanced-learn Version	0.12.4

Exploratory Data Analysis

Extreme Data Imbalance



Label	Count
5	4479
10	1081
6	553
8	516
12	457
---	---
0	18
2	7
22	7
1	7
16	6

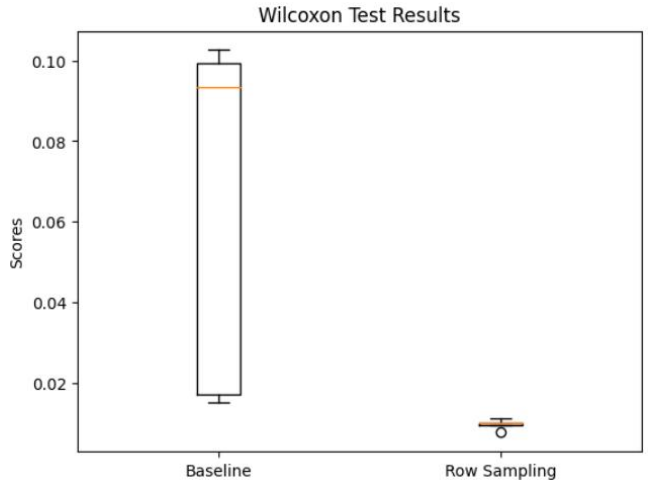
Distribution of Labels

- The majority class represents **44.8%**
- The top-5 frequent classes represent **70.9%**
- The rarest class contains only **6 samples**, accounting for just **0.06%**

Solution: Over & Under Sampling

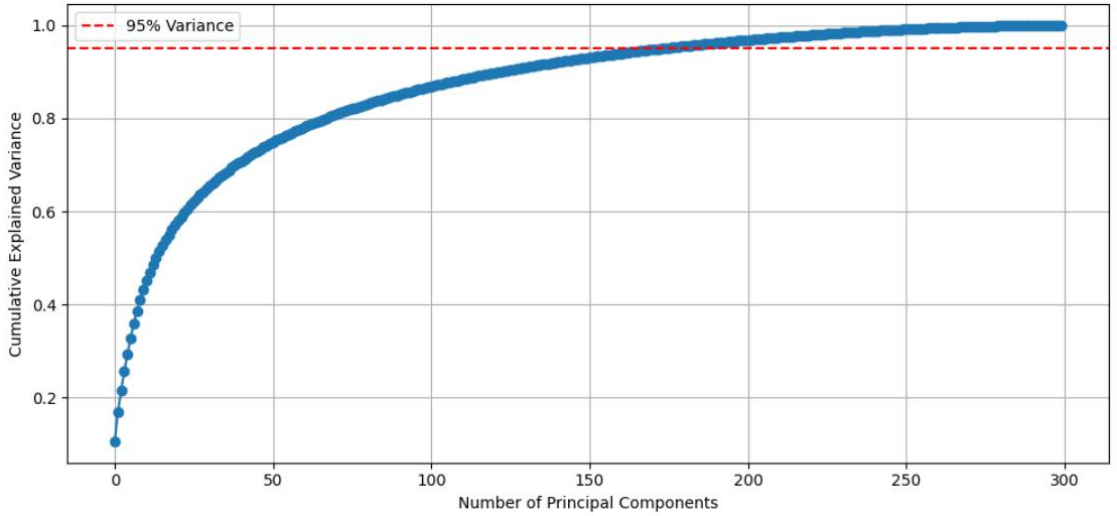
- Random Over & Under Sampling
- **SMOTE + Random Under Sampling + ENN (best performance)**
- ADASYN + TomeLinks

Wilcoxon test p-value: 0.0625



Exploratory Data Analysis

Feature Redundancy



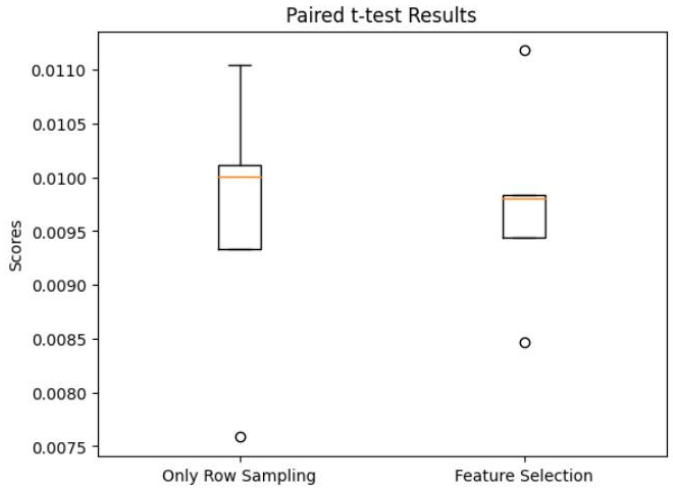
PCA Cumulative Explained Variance

- The first **170 principal components** capture over **95%** of the total variance

Solution: GDBT Model & Feature Selection

- Tree-based models, automatic feature selection by **selecting the most informative features**
- Additional feature selection based on importance/**SHAP (best performance)**

Paired t-test p-value: 0.6527



Other Objectives & Strategies

Bias Reduction

- **Solution:** GDBT Models (XGBoost, LightGBM, CatBoost) with Bayesian hyperparameter tuning

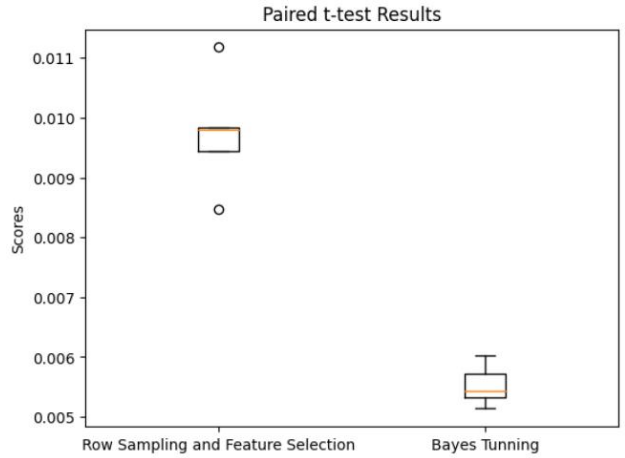
Generalization

- **Solution:** Ensemble learning (k-fold inference) using logistic regression as meta learner

Overall Performance

- **Solution:** Grid search for optimal ratios for over & under sampling, as well as feature selection

Paired t-test p-value: 0.0001



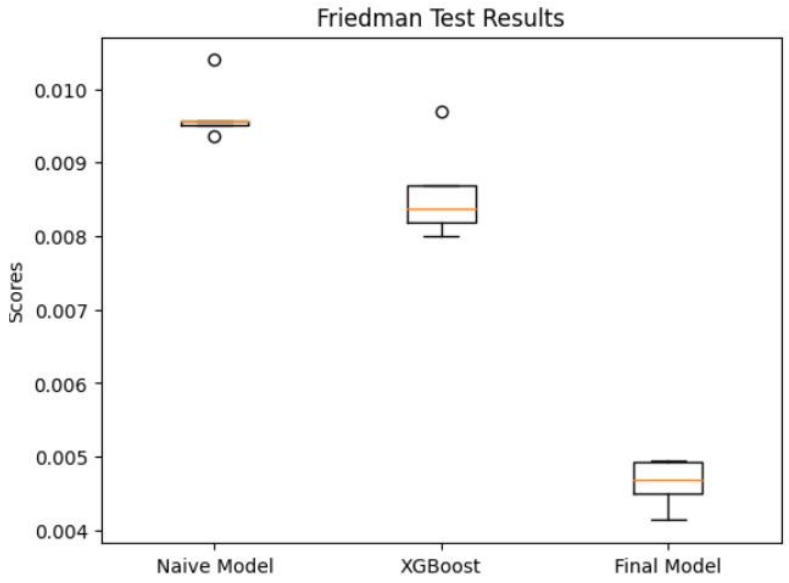
Ensemble Strategy	Score
Soft (Average) Voting	0.005126
Weighted Voting	0.01467
Stacking (Logistic Regression)	0.007238
Stacking (Logistic Regression-L2Norm)	0.005098
Stacking (Random Forest)	0.006608
Stacking (XGBoost)	0.007377

Result

5-Fold Cross Validation Average Result： **0.0046**

- Baseline naive model: 0.0097
- Improvement over 50%

Fold	Weighted Cross-Entropy Loss
1	0.0046828424245336 2,
2	0.0049301640696826 67
3	0.0049566935883394 14
4	0.0041386306294279 97
5	0.0045093451974166 1

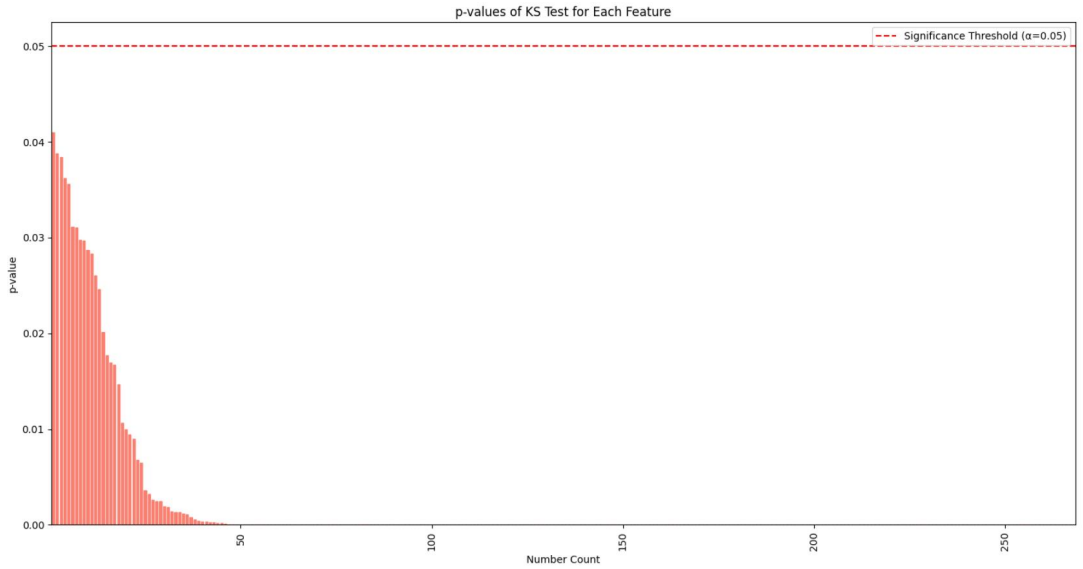


Distribution Shift

Covariate Shift

Experiment: Kolmogorov–Smirnov (K-S) tests on each feature between the training set and Test Set 2

- **Result:** Significant distribution changes in **268/300 features**
- **Conclusion:** **P(X) changes**



Experiment: Make prediction using trained model on 202 known samples

- **Result:** **No significant drop in performance**
- **Conclusion:** **P(y|X) remains the same**

Conclusion: **Covariate shift occurs**, **but no concept drift**

Solution: Retrain model with sample weights

Distribution Shift

Label Shift

Experiment: Kolmogorov–Smirnov (K-S) tests on each feature between the training set and Test Set 2

- **Result:** **Significant distribution change**
- **Conclusion:** **P(y) changes**

Experiment: Kolmogorov–Smirnov (K-S) tests on each feature in each category

- **Result:** **No significant change in distribution**
- **Conclusion:** **P(X|y) remains the same**

Conclusion: **Label shift occurs**

Solution: Bayes posterior correction

Class	Changed Feature Count	Proportion (%)
12	15	5
25	13	4.3
9	7	2.3
27	25	8.3
14	5	1.6
24	10	3.3
8	7	2.3
5	9	3.0
4	10	3.3
17	14	4.7
11	39	13.0
6	22	7.3
23	4	1.3
21	12	4.0
13	7	2.3
7	17	5.7
26	14	4.7
19	16	5.3
10	15	5.0
20	18	6.0
3	23	7.7
18	14	4.7

Discussion

Pros

- Select GDBT models to overcome extreme imbalance
- Perform k-fold inference strategy to prevent overfitting
- Conduct grid search for sampling and feature selection ratios to optimize performance
- Evaluate models through strict statistical test
- Thoroughly detect and address distribution shift based on its formal definitions

Cons & Further Works

- Resource demands incurred by GBDT models
- Try other strategies for handling imbalanced classification include one-vs-all modeling
- Perform grid search with smaller step sizes
- Conduct Further exploration on recent strategies for combating distribution shift

Thank you!