Assignment 2 : Boy or Girl

112423004 CHUN-HUA, SHIH | 112423066 CHIA-HENG, KUO | 112423061 TING-LIN, KUO Group Name:Wikikepia

Introduction

The following describes our experimental design process. In this experiment, we will examine in detail how we solve major problems in data processing, namely: missing values, imbalanced datasets, character selection, and extraction. We will focus on preprocessing, model training, and text-processing techniques to focus on the right data. We will show the methods we have chosen and the reasons behind them, and explore the improvements we have implemented and how these improvements help us to predict gender more accurately.

Preprocessing

Missing value imputation

- Discrete data: We use the mode to fill missing values. There are only 3 operating systems in this data set.
- Continuous data: we chose to fill in the median to avoid too many extreme outliers affecting the entire data.

Detecting and Treating Outliers

- We opted for manually setting upper and lower bounds. Any record outside this range is assigned a value of 0 in the code.

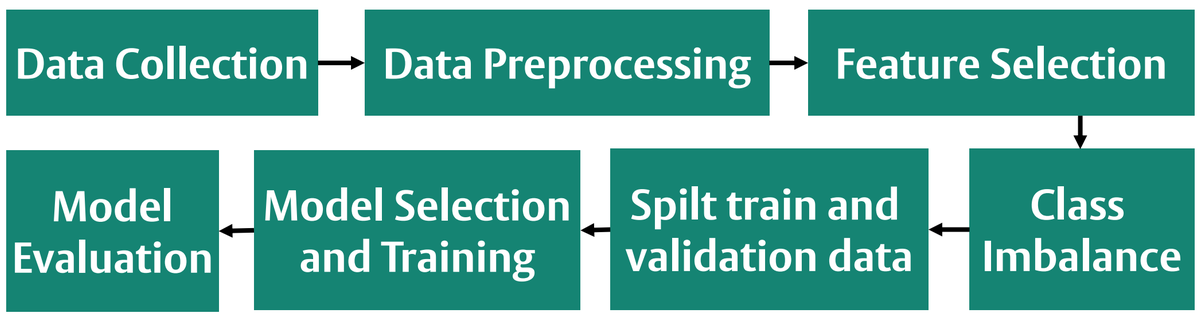
NLP and Feature selection

- Text preprocessing: Our process involves lowercase, handling contractions, removing punctuation, tokenizing words, eliminating stopwords, and lemmatizing words.
- TF-IDF transformation: We Initially extracts nearly 300 terms. Next, we count occurrences, selecting those appearing at least three times. Finally, We applies chi-square test, conducting supervised training with target prediction (gender field), and extracts top 20 features.

Dealing with Imbalanced data: SMOTE

- Dataset Characteristics: original data has a ratio of 316:107 for men and women, which is close to 3:1.
- SMOTE interpolates between existing minority class samples, balancing the class distribution and providing the model with more representative samples of the minority class.

Procedure



Model Selection

We experimented with various methods included KNN, Decision Tree, RF, SVM. After many different comparisons and attempts, we found that the performance of Random Forest was relatively stable, so we finally decided to use RF as the group's classifier.

Experiment

	Data Preprocessing			Model Prediction	Performance	
No	Sampling Method	Semantic Analysis?	Outlier Method	Classifier	Private	Public
1	X	X	X	X	0.50505	0.50761
2	SMOTE	X	Maximum of 5 rows	SVM(RBF)	0.80808	0.82741
3	SMOTE	X	Median 1.5*IQR	SVM(RBF)	0.81313	0.87309
4	SMOTE	X	Median	SVM(RBF)	0.39393	0.34517
5	X	V	Median	KNN	0.59595	0.65482
6	SMOTE	X	Median + setting up/low bound	Random Forest	0.84343	0.89847
7	SMOTE	V	Median + setting up/low bound	Random Forest	0.83838	0.90355
8	SMOTE	V	Median + setting up/low bound	SVM(RBF)	0.55555	0.51776
9	X	V (Total 28 features)	Median + setting up/low bound	Random Forest	0.84343	0.87817
10	X	V (Total 54 features)	Median + setting up/low bound	Random Forest	0.83333	0.86802
11	SMOTE	V	Median + setting up/low bound	Random Forest	0.87373	0.91878

Results

Our best model‘s performance metrics and prediction results are as follows:

	BOY	GIRL
Precision	0.8586956521739131	0.8991596638655462
Recall	0.9518072289156626	0.7328767123287672
F1-score	0.9028571428571429	0.8075471698113209
Total Accuracy	0.8708860759493671	
ROC_AUC	0.8423419706222149	
PRAUC	0.6048597752353954	

Discussion

- In our study, we tackled various data challenges such as imbalance, missing values, and outliers. Median imputation effectively handled missing values, while manual outlier handling ensured data consistency.
- In text mining, TF-IDF and semantic analysis improved gender inference from self-introduction texts.
- During model evaluation, the random forest model with SMOTE emerged as the most effective, demonstrating high accuracy and generalization. Conversely, linear SVM struggled with gender distinction, leading us to exclude it.
- Overall, we address data challenges and leveraging techniques like SMOTE and semantic analysis enhanced predictive accuracy and informed our selection of the most suitable classification approach.