CS4622 - Machine Learning
Lab 01 - Feature Engineering
Report

P. L. Nimantha Dilshan Cooray

190111B

Dataset Exploration

Two datasets were given for this lab assignment.

1. train.csv – 28520 rows
2. valid.csv – 750 rows

Later a test dataset (750 rows) was also given to evaluate the feature engineering.

There are 256 features and 4 labels in all datasets.

Handling Missing Values in `label_2` (Speaker Age) column

As given in the description, there were missing values in the `label_2` (Speaker Age). Steps were performed to find how many null values are there in label_2.

	Missing Value Count	Missing Value Percentage
Train Dataset	480	1.683
Validation Dataset	14	1.867
Test Dataset	6	0.8

Table 1 - Missing Values in label_2 (Speaker Age)

Since the missing value count and percentage are small, those rows were dropped from the datasets that are used to predict the `label_2`.

Handling Imbalanced `label_4` (Speaker Accent) column

The distribution of values in the label_4 column was not equally distributed.

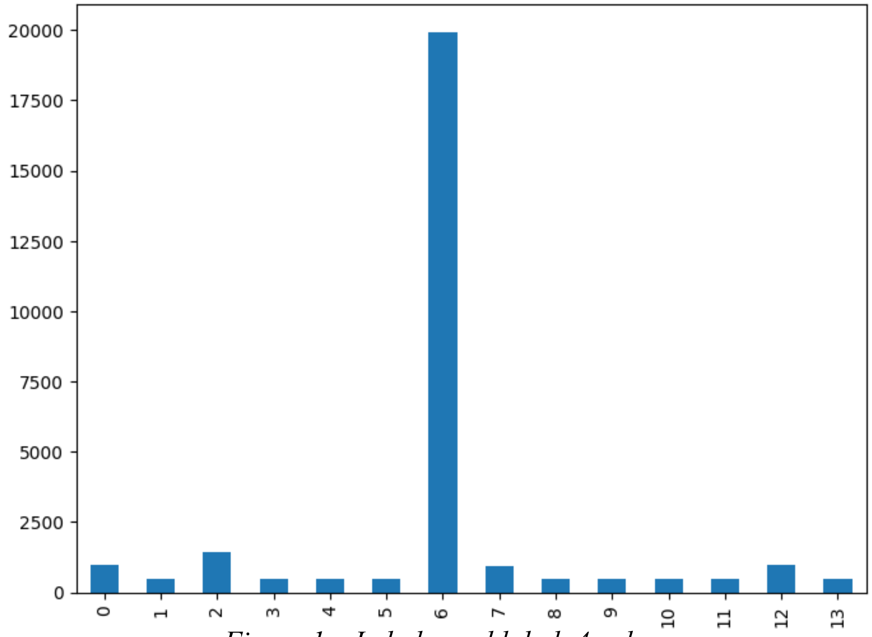

Figure 1 – Imbalanced labal_4 column

As a workaround, tried to under-sample the data. This means we are deleting the rows of the majority class from the dataset such that all the classes will have an equal distribution. This approach didn't work well because the accuracy of machine learning models was low when we resampled the data. The below table shows the accuracies of models trained on under-sampled data for predicting `label_4`.

	SVM		Random Forest	
	Validation	Test	Validation	Test
`label_4`	0.779	0.78	0.413	0.411

Table 2 – Accuracy after undersampling label_4

Baseline Models

Creating a baseline is essential to compare results after feature engineering. Multiple models were trained and evaluated for each label.

For each label, baseline models were trained on all 256 features. Before giving the input to the models, all the features were standardized. Note that rows with missing values for `label_2` were dropped in a previous step.

Since label_1, label_3, and label_4 are categorical, classification models SVM and Random Forest were used. For label_2 (Speaker Age), a regression method offered by XGBoost is used.

	SVM (Classification)		Random Forest (Classification)		XGBoost (Regression) (RMSE)	
	Validation	Test	Validation	Test	Validation	Test
`label_1` (Speaker ID)	0.991	0.989	0.967	0.968	-	-
`label_2` (Speaker Age)	-	-	-	-	3.29	3.143
`label_3` (Speaker Gender)	0.995	1	0.995	0.997	-	-
`label_4` (Speaker Accent)	0.937	0.933	0.844	0.857	-	-

Table 3 – Accuracy of the baseline models

Feature Engineering

Principal Component Analysis

Principal Component Analysis (PCA) is used to reduce the number of dimensions in the input. The PCA class of the sklearn library was used with 0.95 as the `n_components` parameter and `‘full’` as the `svd_solver` parameter. This means PCA is done in a way to select components that can explain 0.95 of the total variance.

By doing PCA, the original 256 features were reduced to 67 features (components). The accuracy of the reduced features set is shown below.

	SVM (Classification)		Random Forest (Classification)		XGBoost (Regression) (RMSE)	
	Validation	Test	Validation	Test	Validation	Test
`label_1` (Speaker ID)	0.981	0.973	0.967	0.975	-	-
`label_2` (Speaker Age)	-	-	-	-	3.675	3.403
`label_3` (Speaker Gender)	0.999	0.997	0.999	0.992	-	-
`label_4` (Speaker Accent)	0.908	0.924	0.821	0.831	-	-

Table 4 – Accuracy of the models trained after applying PCA

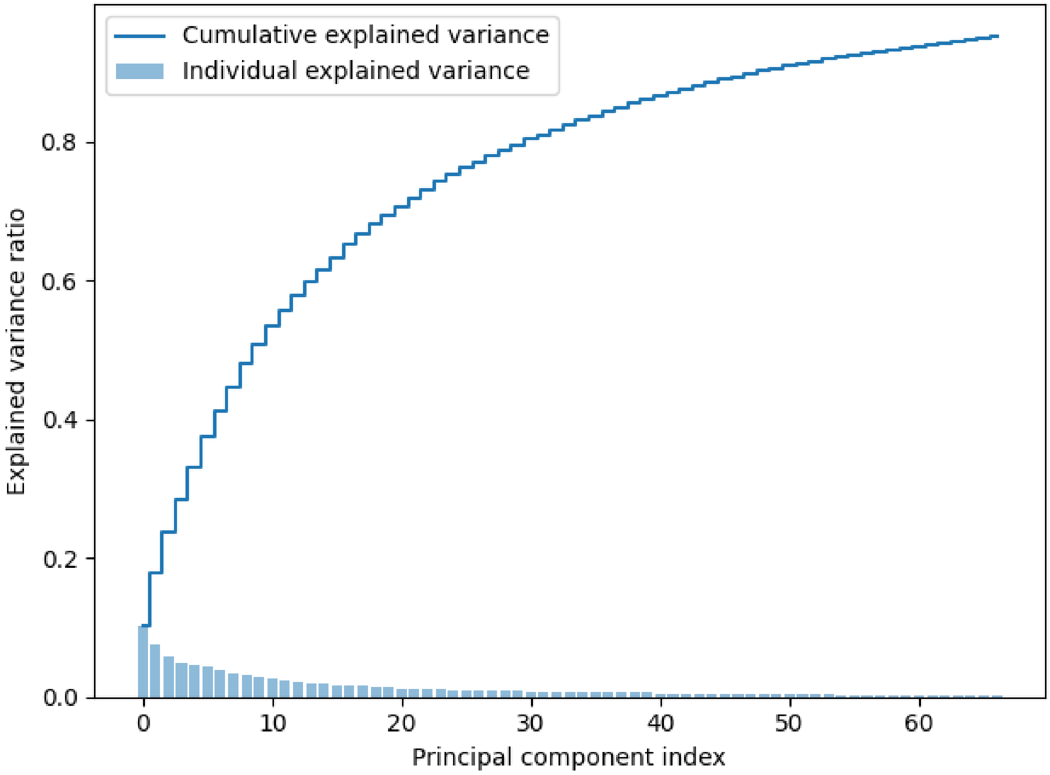

Figure 2 – Explained variance and Cumulative explained variance chart

ANOVA F-Test

The ANOVA F-test was done to see if it could produce a reduced set of features that can be used to achieve higher accuracy. Using sklearn library classes, a feature set of 67 was obtained. For this method, we need to give the number of features(k) we need. The top k features with the highest F-score will be selected. The number 67 was chosen as the number of features to see if this method can perform better than PCA.

The accuracy of the models trained using the feature set obtained by this method was relatively low compared to the PCA approach.

	SVM (Classification)		Random Forest (Classification)		XGBoost (Regression) (RMSE)	
	Validation	Test	Validation	Test	Validation	Test
`label_1` (Speaker ID)	0.957	0.956	0.947	0.928	-	-
`label_2` (Speaker Age)	-	-	-	-	7.282	6.64
`label_3` (Speaker Gender)	0.997	0.997	0.993	0.993	-	-
`label_4` (Speaker Accent)	0.895	0.913	0.84	0.855	-	-

Table 5 – Accuracy of the models trained after applying F-test

Final Feature Selection

From the above approaches, the feature set derived by PCA is submitted as the final feature set. That feature set was used to train an SVM model and predict all four labels. The results were uploaded as CSV files to the submission links.

You can access the python notebook from here.