

README

Predicting if a business in America will survive COVID-19



CS 760

12.07.2020

Mu Cai, Varun Sreenivasan, Niharika Tomar

¹ <https://research.qut.edu.au/ace/2020/04/30/covid19-pandemic-as-external-enabler-of-entrepreneurship/>

This file provides information on how to run the source code for the classification models

LIST OF FILES AND FOLDERS

1. RAW_DATASET

- yelp_academic_dataset_business.json: General business information for the respective businesses.
- yelp_academic_dataset_covid_features.json: Covid specific information for the respective businesses.

2. EXTERNAL_DATA

- Population: A dataset containing population figures relative to an American postcode.
- State_Incomes: A dataset containing distribution of people across income brackets relative to an American postcode. Official: The original income dataset.
 - Official: The original income dataset.
 - Parsed: A parsed version of the original that is used to build our main dataset.

3. PRE-PROCESSING

- JSONCSV.ipynb: Script that calls multiple functions from files listed below to perform pre-processing to create the main dataset from the raw Yelp datasets and the external datasets listed above.
- business_attributes.py: contains functions to parse attributes from Yelp business dataset.
- business_categories.py: contains functions to map the sub categories to the main category.
- covid_feature_extraction.py: contains functions to parse Covid features.

- `income_extraction.py`: contains functions to parse Income dataset to help assign median income category for a postcode.
- `population_extraction.py`: contains functions to parse the population dataset to obtain population relative to a postcode.

4. PROCESSED_DATASET

- `Dataset.csv`: The main dataset generated after pre-processing.

5. CROSS_VALIDATION

- `cross_validation.py`: Validation pipeline that performs 10-Fold CV. On each fold, SMOTE is applied to the training set. Balanced accuracy is used to evaluate performance on validation sets.
- `cross_validation_normalization.py`: Same as `cross_validation.py` but Min-Max normalization is also applied in the pipeline.

6. CLASSIFICATION_MODELS

- `KNN.ipynb`
- `Logistic-Regression.ipynb`
- `Random-Forest.ipynb`
- `Neural-Nets.ipynb`
- `Naive-Bayes.ipynb`
- `SVM.ipynb`

7. FEATURE_IMPORTANCE

- `Feature-Importance.ipynb`: Determines the top 10 important features using Permutation Importance algorithm on Random Forest.

8. TEX_FILES











- main.tex: The main body of our tex file.
- signi_feature.png: The figure needed in our main.tex, which depicts the significant features.

9. REPORT_PDF

- cs760_report_Mu_Cai_Varun_Sreenivasan_Niharika_Tomar.pdf: The PDF report of our project.

EXECUTING CODE FOR CLASSIFIERS

- Make sure Dataset.csv, cross_validation.py, cross_validation_normalization.py and the ipynb file corresponding to the classifier are in the same folder like in the image below

 cross_validation	12/5/2020 9:38 PM	Python File	2 KB
 cross_validation_normalization	12/3/2020 2:10 AM	Python File	2 KB
 Dataset	11/23/2020 5:16 AM	Microsoft Excel Com...	17,949 KB
 Feature-Importance.ipynb	12/7/2020 4:32 AM	IPYNB File	47 KB
 KNN.ipynb	12/7/2020 3:38 AM	IPYNB File	7 KB
 Logistic-Regression.ipynb	12/7/2020 2:01 AM	IPYNB File	7 KB
 Naive-Bayes.ipynb	12/7/2020 4:23 AM	IPYNB File	9 KB
 Neural-Nets.ipynb	12/7/2020 3:40 AM	IPYNB File	32 KB
 Random-Forest.ipynb	12/7/2020 2:24 AM	IPYNB File	7 KB
 SVM.ipynb	12/7/2020 4:00 AM	IPYNB File	7 KB

EXECUTING CODE FOR FEATURE-IMPORTANCE

- Make sure Dataset.csv and the Feature-Importance ipynb files are in the same folder like in the image

above.

EXECUTING CODE TO GENERATE PROCESSED DATASET

- Make sure yelp_academic_dataset.json, yelp_academic_dataset_covid_features.json, Population Folder, State_Incomes Folder along with the python files business_attributes.py, business_categories.py, covid_feature_extraction.py, income_extaction.py, population_extraction.py and the ipynb file JSONCSV.ipynb are in the same folder like the image below. Run the cells in JSONCSV.ipynb

	Population	12/9/2020 4:07 AM	File folder	
	State_Incomes	12/9/2020 4:07 AM	File folder	
	business_attributes	12/9/2020 4:03 AM	Python File	9 KB
	business_categories	12/9/2020 4:03 AM	Python File	44 KB
	covid_feature_extraction	12/9/2020 4:03 AM	Python File	2 KB
	Dataset	12/9/2020 4:08 AM	Microsoft Excel Com...	17,949 KB
	income_extraction	12/9/2020 4:03 AM	Python File	16 KB
	JSONCSV.ipynb	12/9/2020 4:09 AM	IPYNB File	14 KB
	population_extraction	12/9/2020 4:03 AM	Python File	23 KB
	yelp_academic_dataset_business	12/9/2020 4:03 AM	JSON File	149,316 KB
	yelp_academic_dataset_covid_features	12/9/2020 4:03 AM	JSON File	63,316 KB