AirBnB Destinations

Project Design

This project uses the AirBnB Challenge dataset from a 2015 Kaggle challenge. The target of this challenge is `country_destination`, the country a new AirBnB traveled in their first booking. There are 12 classes to classify on, of which I drop one, `NDF` , for users that haven't booked a first trip.

I built three subset models, which I rank from simplest to most complex here:

Model	Number of Targets
USA vs France	2
USA vs Not-USA	2
Multiclass	11

Tools

Pandas, Numpy, SKLearn, imblearn, scipy

Sklearn submodules: RandomizedSearchCV, LogisticRegression, KNN, SVC, preprocessing, pipeline, model_selection.

Data

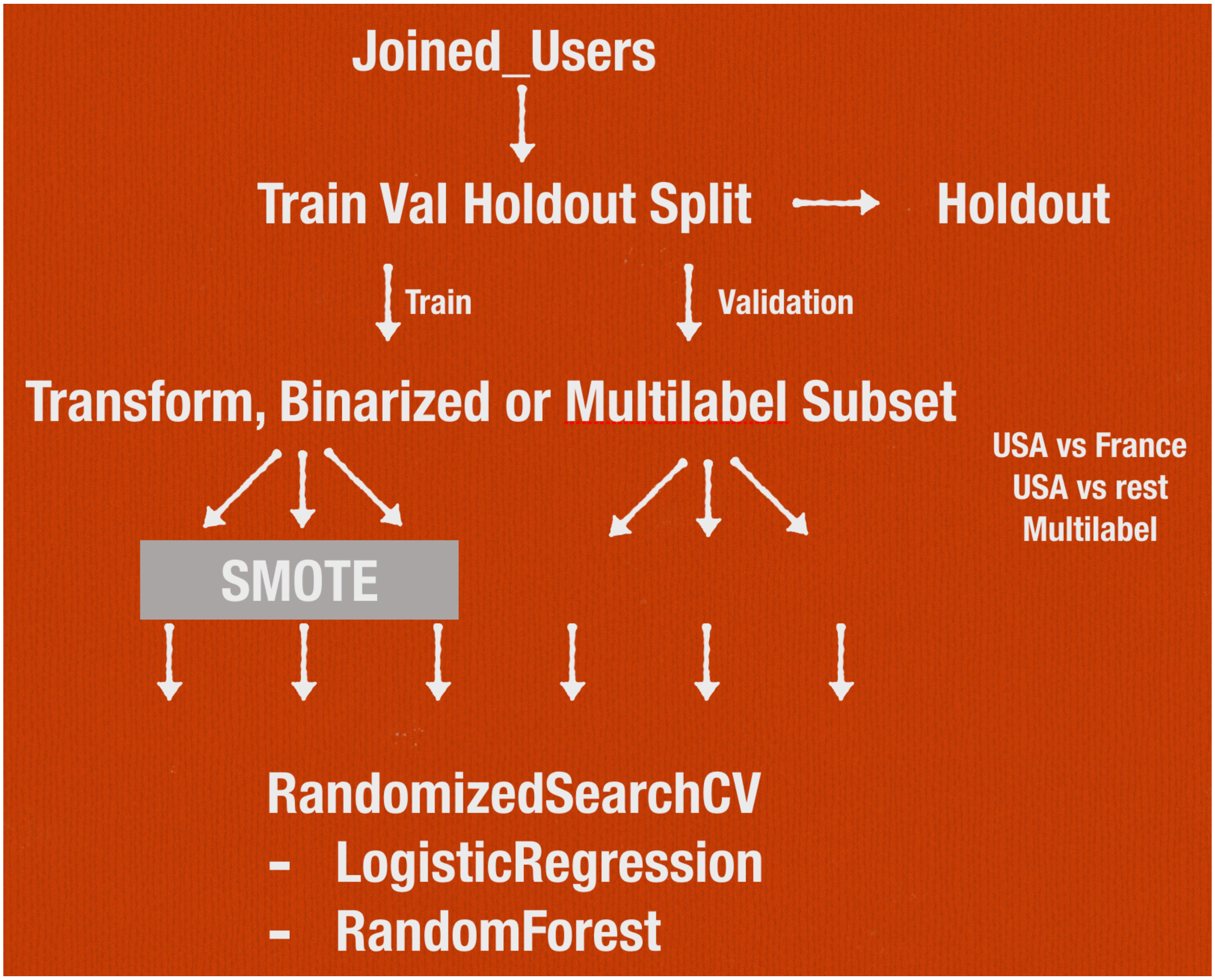
The data are stored in four separate .csv tables:

- Age Gender Brackets (12 kb)
- Countries (<1 kb)
- Sessions (632 MB)
- Train Users (25 MB)

Table	List of columns	Size
`age_gender_bkts`	(`age_bucket`, `country_destination`, `gender`, `population_in_thousands`, `year`)	12 kb
`countries`	(`country_destination`, `lat_destination`,`lng_destination`, `distance_km`, `destination_km2`, `destination_language`, `language_levenshtein_distance`)	1 kb
`sessions`	(`user_id`, `action`, `action_type`, `action_detail`, `device_type`, `secs_elapsed`)	632 MB
`train_users`	(`id`, `date_account_created`, `timestamp_first_active`, `date_first_booking`, `gender`, `age`, `signup_method`, `signup_flow`, `language`, `affiliate_channel`, `affiliate_provider`, `first_affiliate_tracked`, `signup_app`, `first_device_type`, `first_browser`)	25 MB

Algorithms

I built extensive data cleaning and feature engineering code code in `airbnb.py`, which results in a well-formated, feature-engineered DataFrame `Joined_Users`. This DataFrame then was split, transformed into the three model subsets, upsampled via `SMOTE` , and optimized via a hyperparameter search using `RandomizedSearchCV`. The scheme is capture below:



Joined_Users

Train Val Holdout Split

Holdout

Train

Validation

Transform, Binarized or Multilabel Subset

USA vs France
USA vs rest
Multilabel

SMOTE

RandomizedSearchCV
- LogisticRegression
- RandomForest

What to do different next time

This was a challenging project, in large part because it was difficult to form an intuition around the data and how it best relates to the target. Next time, I'd try to get a better strategy down on data cleaning and feature engineering and do so sooner rather than later.