Predicting Star Ratings

PROJECT OVERVIEW:

The goal of this project is to predict a new venue's popularity from information available when the venue opens. This project is preformed by machine learning from a data set of venue popularities provided by Yelp.

The data set contains meta data about the venue (where it is located, the type of food served, etc.). It also contains a star rating.


```
[{'business_id': 'vcNAWiLM4dR7D2nwwJ7nCA',
  'full_address': '4840 E Indian School Rd\nSte 101\nPhoenix, AZ 85018',
  'hours': {'Tuesday': {'close': '17:00', 'open': '08:00'},
   'Friday': {'close': '17:00', 'open': '08:00'},
   'Monday': {'close': '17:00', 'open': '08:00'},
   'Wednesday': {'close': '17:00', 'open': '08:00'},
   'Thursday': {'close': '17:00', 'open': '08:00'}},
  'open': True,
  'categories': ['Doctors', 'Health & Medical'],
  'city': 'Phoenix',
  'review_count': 7,
  'name': 'Eric Goldberg, MD',
  'neighborhoods': [],
  'longitude': -111.983758,
  'state': 'AZ',
  'stars': 3.5,
  'latitude': 33.499313,
  'attributes': {'By Appointment Only': True},
  'type': 'business'},
 {'business_id': 'JwUE5GmEO-sH1FuwJgKBlQ',
  'full_address': '6162 US Highway 51\nDe Forest, WI 53532',
  'hours': {},
  'open': True,
  'categories': ['Restaurants'],
  'city': 'De Forest',
  'review_count': 26,
  'name': 'Pine Cone Restaurant',
  'neighborhoods': [],
  'longitude': -89.335844,
  'state': 'WI',
  'stars': 4.0,
  'latitude': 43.238893,
  'attributes': {'Take-out': True,
   'Good For': {'dessert': False,
    'latenight': False,
    'lunch': True,
    'dinner': False,
    'breakfast': False,
    'brunch': False},
   'Caters': False,
   'Noise Level': 'average',
```

PROJECT OUTLINE:

The prediction consists four models:

- **city_model**
  - A custom estimator that based solely on the city of a venue (average star of a city)
- **lat_long_model**
  - City-based model might not be sufficiently fine-grained
  - The latitude and longitude of a venue are used as features to understand neighborhood dynamics
  - A ColumnSelectTransformer is used to transform latitude and longitude values to an array containing selected keys of feature matrix
- **category_model**
  - Besides location, the venue's category is also predictive. An estimator that considers the categories is built.
  - Use one-hot encoding (DictVectorizer) to deal with categorical features.
- **attribute_model**
  - There is more information in the attributes for each venue. An estimator based on these attributes (e.g., Attire, Ambiance, Good for, Noise level, Caters…)
  - Firstly, flatten the nested dictionary to a single level, then encode them with one-hot encoding

RESULTS:

- Combine all the models together

```
from sklearn.pipeline import FeatureUnion

union = FeatureUnion([('city_model', EstimatorTransformer(city_est)),
  ('lat_long_model', EstimatorTransformer(scaled_nearest_neighbors_cv)),
  ('category_model', EstimatorTransformer(category_est_cv)),
  ('attribute_model', EstimatorTransformer(attributes_est_tot))

        # FeatureUnions use the same syntax as Pipelines
    ])
```

- Test case