The perfect night out

IBM Applied Data Science Specialization Capstone Report

By Mike Levenson

PROBLEM:

People (both tourists and residents alike) can become overwhelmed when planning a night out in a large city such as New York due to the seemingly endless amount of choices. This situation can actually become stressful and lead to less than optimal experiences. The notebook will focus on building a classification model that analyzes nightlife data to recommend optimal bars according to the type of desired experience:

- lively venues (coded: L)
- relaxed venues("chill" coded: C)
- romantic venues(coded: R).

Location: Williamsburg Neighborhood, Brooklyn, NY

- Data is extracted by calling the api for 50 nightlife venues in the neighborhood and then details for each venue.
  - target attributes for the model are filtered and extracted from the details data set which are a combination of
    - categories
    - venue types
    - price

METHODOLOGY:

This problem calls for a heuristic approach because results and target attributes can differ.

- The data: 50 venues
  - Split with 35 venues to train and 15 venues to test.
- The model: multiclass Support Vector Machine model was employed to discover labels of nightlife venues. Parameters were tweaked to support multi-class decisions and output optimal prediction given the circumstances..
  - Decision function shape was put to 'ovo' or One versus One.
  - The gamma set to scale.
  - Class-weight 'balanced' due to an imbalance in training set labels.

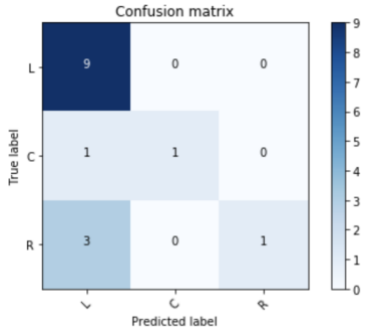
```
In [38]: from sklearn import svm
         SVM = svm.SVC(kernel='rbf',decision_function_shape='ovo', gamma='scale', class_weight='balanced')
         SVM.fit(X_train, y_train)

Out[38]: SVC(C=1.0, cache_size=200, class_weight='balanced', coef0=0.0,
           decision_function_shape='ovo', degree=3, gamma='scale', kernel='rbf',
           max_iter=-1, probability=False, random_state=None, shrinking=True,
           tol=0.001, verbose=False)
```

RESULTS:

After optimizing the model to the best of my abilities, the model classified

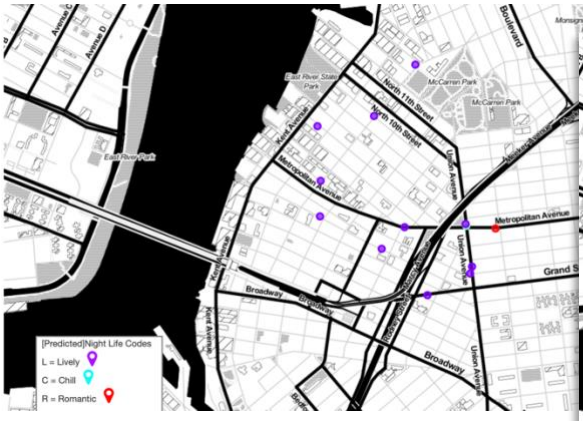
- Lively venues with an F1 score of 0.82
- Relaxed(Chill) venues with an F1 score of 0.67
- Romantic venues with an F1 score of 0.40 while being confused with Lively venues.



	precision	recall	f1-score	support
C	1.00	0.50	0.67	2
L	0.69	1.00	0.82	9
R	1.00	0.25	0.40	4
micro avg	0.73	0.73	0.73	15
macro avg	0.90	0.58	0.63	15
weighted avg	0.82	0.73	0.69	15

```
Confusion matrix, without normalization
[[9 0 0]
 [1 1 0]
 [3 0 1]]
```

Predicted Nightlife Labels



Actual Nightlife Labels

Discussion & Conclusion

There were multiple issues during this project:

1. Foursquare API
    A. detail responses were difficult to work with due to heavily layered JSON outputs
    B. Venue detail output unformity is lacking and missing in some cases which creates a data wrangling challenge.
2. Due to a heavy imbalance in class labels during training of the model, confusion of labels tended to label venue types as Lively (majority of venues were labeled as lively in training.
    A. future approaches will need to take a large set of each classification and train on them separately in a binary approach (ex: 30 Lively venues labeled Lively or not, 30 Chill venues labeled Chill or not etc...)
3. Another future approach would be to text mine user responses/tips/labels and establish categories from uniformly assigned categories based on user responses.

In conclusion, this notebook is a solid attempt to classify nightlife as it offers a legitimate alternative to recommender systems where user data may be sparse in the beginning. There are many approaches that can be taken to vastly improve accuracy of this model.