California Restaurant "Likes" Prediction Using Foursquare API and Machine Learning

By : Khalid,Meguenni

Capstone Project

IBM Data Science Professional Certificate

1. Introduction

1. Introduction

- California boasts an incredibly diverse collection of restaurants catering to different palettes and appetites. A large part of marketing for a modern restaurant (or any company) is the number of "likes" on Social Media.
- For a new business owner (or existing company) to open a new restaurant in California, knowing ahead of time the potential social media image they can have would provide an excellent solution to the ever present business problem of uncertainty.
- We can mitigate this uncertainty through leveraging data gathered from FourSquare's API.
- The question we will try to address is, how accurately can we predict the amount of "likes" a new restaurant opening in this region can expect to have based on the type of cuisine it will serve and which city in California it will open in.

2. Data

2.1 Data Scraping and Cleaning

2.1 Data Scraping and Cleaning

- We will first retrieve the geographical coordinates of the three cities (San Francisco, Los Angeles, and San Diego) to represent California.

- Then, we will leverage the FourSquare API to obtain URLs that lead to the raw data in JSON form.

- It is important to note that the extracts are not of every restaurant in those cities but rather all of the venues within a 1000KM range of the geographical coordinates that geolocator was able to provide. As such the data will have to be cleaned before pulling "likes" data.

2.2 Data Preparation

2.2 Data Preparation

- The data still needs some more processing before it is suitable for model training and testing.

- The "categories" column contains too many different types of cuisines to allow a model to yield any meaningful results.

- However, the different types of natural cuisines have natural groupings based on conventionally accepted cultural groupings of cuisine.

2.2 Data Preparation

- As this project will compare both linear and logistic regression, it makes sense to have "likes" as both a continuous and categorical (but ordinal) variable. In the case of turning into a categorical variable, we can bin the data based on percentiles and classify them into these ordinal percentile categories.

- As the last stage of data preparation, it is important to note that the regressors are categorical variables (3 different cities and 6 different categories of cusines). Hence, they require dummy variable encoding for meaningful analysis. We can accomplish this via one-hot encoding.

3. Methodology

3. Methodology

- This project will utilize both linear and logistic regression machine learning methods to train and test the data. Namely, linear regression will be used in an attempt to predict the number of "likes" a new restaurant in this region will have. We will utilize the Sci-Kit Learn Package to run the model.

- We can also utilize logistic regression as a classification method rather than direct prediction of the number of likes. Since the number of "likes" can be binned into different categories based on different percentile bins, it is also potentially possible to see which range of "likes" a new restaurant in this region will have.

- Since the "likes" are binned into multiple (more than 2) categories, the type of logistic regression will be multinomial. Additionally, although the ranges are indeed discrete categories, they are also ordinal in nature. Therefore the logistic regression will need to be specified as being both multinomial and ordinal. This can be done through the Sci-Kit Learn Package as well.

4. Results

4.1 Linear Regression Results

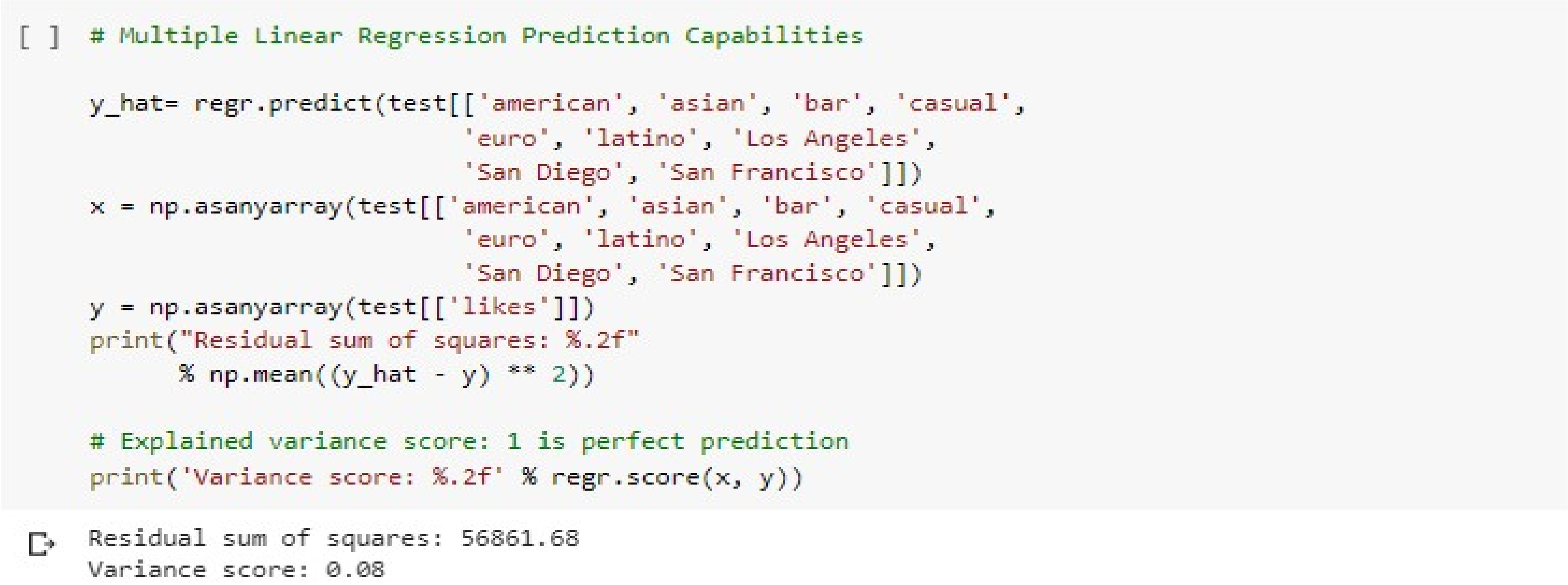
4.1 Linear Regression Results

- A linear regression model was trained on a random subsample of 80% of the sample and then tested on the other 20%.

- To see if this is a reasonable model. the residual sum of squares score and variance score were both calculated (56861,68 and 0.08 respectively).

- Given the low variance score, this is probably not a valid/good way of modelling the data.

4.1 Linear Regression Results



```
# Multiple Linear Regression Prediction Capabilities

y_hat= regr.predict(test[['american', 'asian', 'bar', 'casual',
                         'euro', 'latino', 'Los Angeles',
                         'San Diego', 'San Francisco']])
x = np.asanyarray(test[['american', 'asian', 'bar', 'casual',
                         'euro', 'latino', 'Los Angeles',
                         'San Diego', 'San Francisco']])
y = np.asanyarray(test[['likes']])
print("Residual sum of squares: %.2f"
      % np.mean((y_hat - y) ** 2))

# Explained variance score: 1 is perfect prediction
print('Variance score: %.2f' % regr.score(x, y))
```

```
Residual sum of squares: 56861.68
Variance score: 0.08
```

4.2 Logistic Regression Results

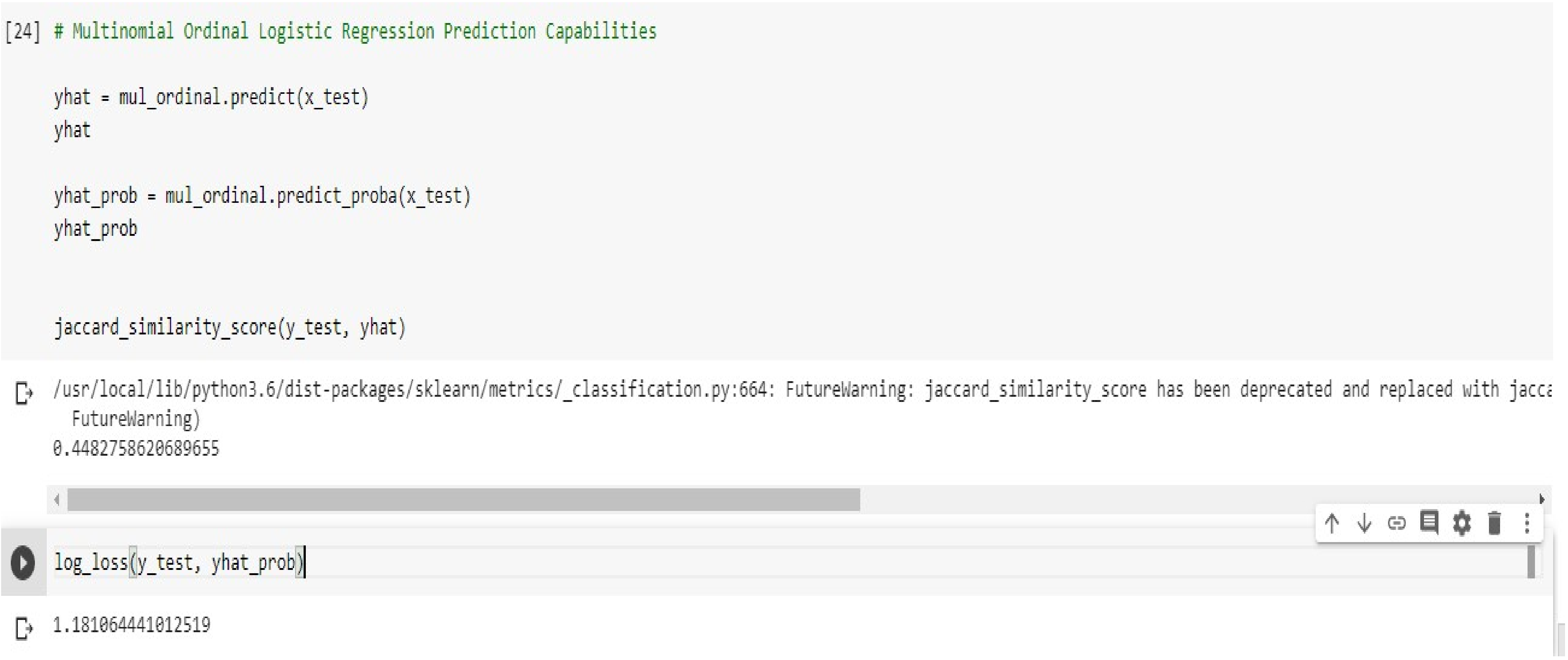
4.2 Logistic Regression Results

- A multinomial ordinal logistic regression model was trained on a random subsample of 80% of the sample and then tested on the other 20%.

- To see if this is a reasonable model, its jacquard similarity score and log-loss were calculated (44% and 1.18 respectively).

- Given the modestly accurate ability of this model, we can also run the model on the full dataset. The coefficients show that opening a restaurant in San Francisco, opening a bar, or serving cuisine that is American or Asian in nature, are associated negatively with "likes."

4.2 Logistic Regression Results



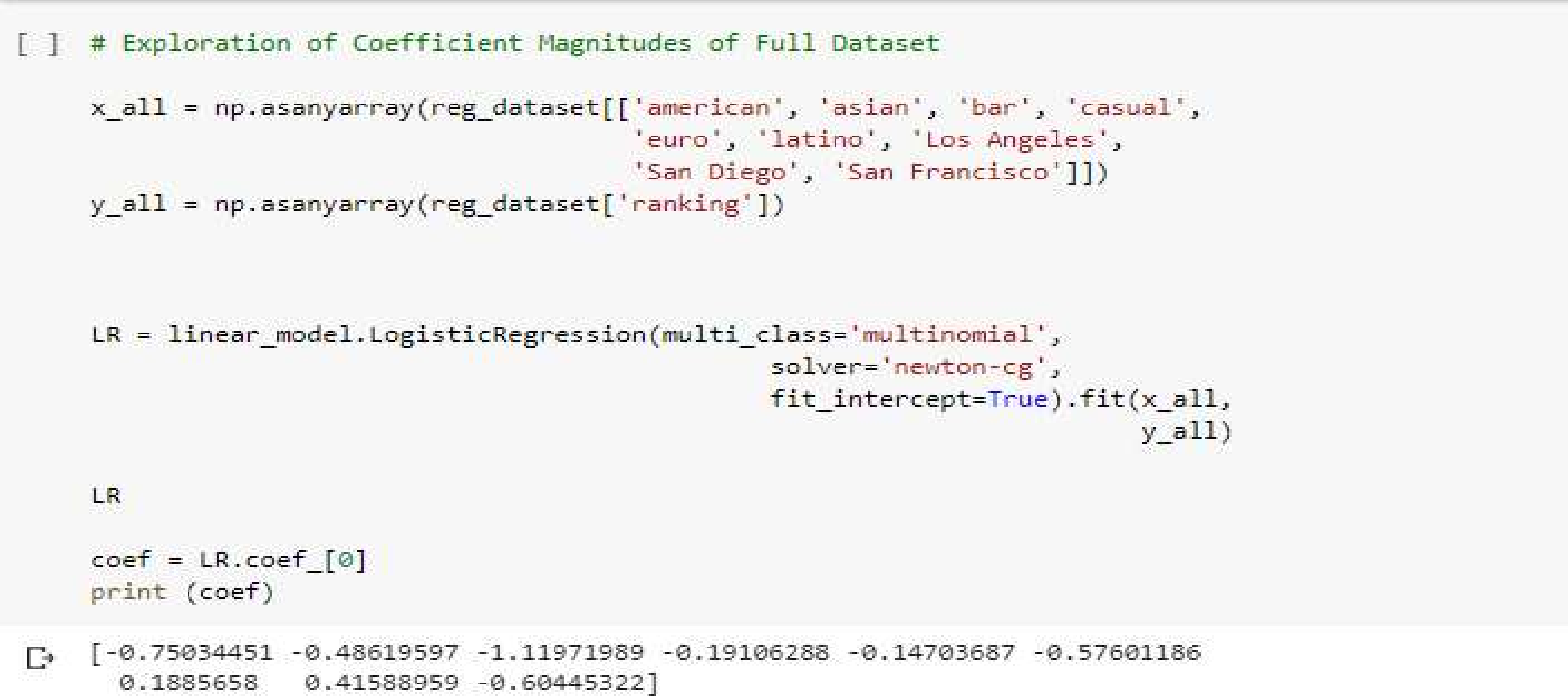
```
[24] # Multinomial Ordinal Logistic Regression Prediction Capabilities

yhat = mul_ordinal.predict(x_test)
yhat

yhat_prob = mul_ordinal.predict_proba(x_test)
yhat_prob



jaccard_similarity_score(y_test, yhat)
```

```
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/_classification.py:664: FutureWarning: jaccard_similarity_score has been deprecated and replaced with jacca
  FutureWarning)
0.4482758620689655
```

```
log_loss(y_test, yhat_prob)
```

```
1.181064441012519
```

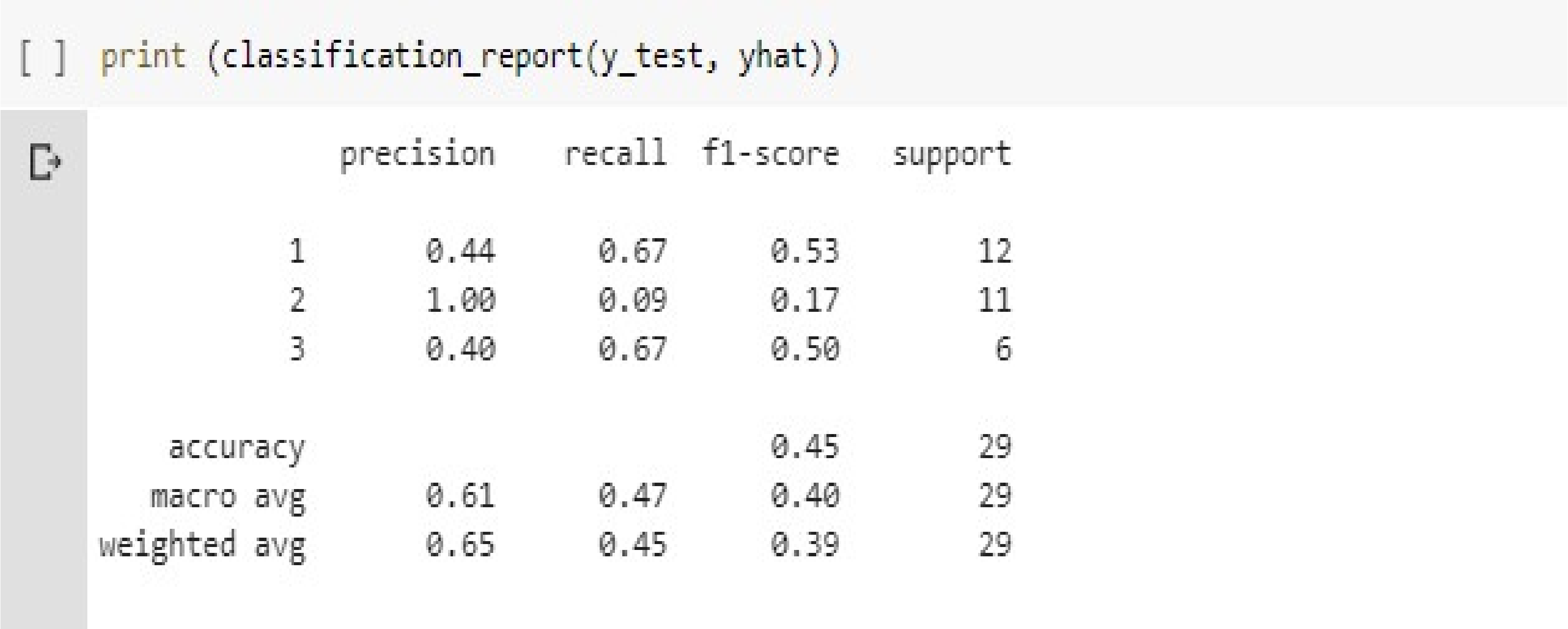
4.2 Logistic Regression Results



```
# Exploration of Coefficient Magnitudes of Full Dataset

x_all = np.asanyarray(reg_dataset[['american', 'asian', 'bar', 'casual',
                                   'euro', 'latino', 'Los Angeles',
                                   'San Diego', 'San Francisco']])
y_all = np.asanyarray(reg_dataset['ranking'])



LR = linear_model.LogisticRegression(multi_class='multinomial',
                                     solver='newton-cg',
                                     fit_intercept=True).fit(x_all,
                                                            y_all)


LR

coef = LR.coef_[0]
print (coef)
```

```
[-0.75034451 -0.48619597 -1.11971989 -0.19106288 -0.14703687 -0.57601186
  0.1885658   0.41588959 -0.60445322]
```

4.2 Logistic Regression Results



```
print (classification_report(y_test, yhat))
```

```
              precision    recall  f1-score   support

           1       0.44      0.67      0.53        12
           2       1.00      0.09      0.17        11
           3       0.40      0.67      0.50         6

    accuracy                           0.45        29
   macro avg       0.61      0.47      0.40        29
weighted avg       0.65      0.45      0.39        29
```

5. Discussion

5. Discussion

- The first thing to note is that given the data, logistic regression presents a better fit for the data over linear regression. Using logistic regression we were able to obtain a Jaccard Similarity Score of 66.66%, which although not perfect, is more reasonable than the low variance score obtained from the linear regression.
- Based on the classification report, its also clear the model is better at predicting if a restaurant will fall into the best or worst percentile of likes.
- In terms of strategy, it can be seen that opening a restaurant in San Francisco, opening a bar, or serving cuisine that is american or asian in nature, are associated negatively with "likes."
- This suggests that the business opportunity should be opening a restaurant in either Los Angeles or San Diego, with a cuisine that is European, Latino, or casual in nature would be the best approach for maximizing likes.

6. Conclusion

6. Conclusion

- In conclusion, after analyzing restaurant "likes" in California from 300 restaurants, we can conclude that the approach to best take is to open a restaurant that is either European, Latino, or casual and that opening the venue in either Los Angeles or San Diego rather than San Francisco.
- Additionally, the predictive capabilities of the logistic regression prediction model are most accurate for classifying whether a restaurant will fall in either the best or worst classes when the data is binned into 3 classes.