Machine Learning Assignment

ACADEMIC HONESTY

As usual, the standard honor code and academic honesty policy applies. We will be using automated **plagiarism detection** software to ensure that only original work is given credit. Submissions isomorphic to (1) those that exist anywhere online, (2) those submitted by your classmates, or (3) those submitted by students in prior semesters, will be detected and considered plagiarism.

INSTRUCTIONS

This assignment tests your ability to create a simple regression solution to a classification problem. You will use a dataset of complaints from 311, the telephone number used for complaints in New York City. You will do a regression that will predict the class of complaints: either high or low processing time complaints.

There are two questions in this assignment.

The file nyc_data.csv (in Vocareum) contains an extract of the 311 complaints data downloaded from the 311 website. Goal: use this data to do a regression analysis that focuses on predicting high and low processing time complaints.

Question 1: Preprocessing

In this question, you will prepare the data before building your regression model. After preparing the data, you will save it as a CSV.

Follow these steps to prepare the data:

1. Read the data with the pandas read_csv function
2. Create a column 'month' that extracts the month number from the 'Created Date' column
3. Create a column 'hour' that extracts the hour of the day from the 'Created Date' column
4. Create a column 'weekday' that extracts the day of the week the complaint was created. You will need to use the datetime library for this.
5. Create a column 'agency_num' that uses a number for each agency. Use .unique() to extract a list of agencies and use the place value of an agency in this list as the agency number). Helping code for this question has been added to the Starter Code of this assignment on Vocareum.
6. Create a column 'borough_num' that uses a number for each borough (Manhattan -> 1, Brooklyn -> 2, Queens -> 3, Staten Island -> 4, Bronx -> 5, Other -> 0). You should count '07 Bronx' as 'Bronx' and *not* as unknown. For example, you can do a test like that: 'Bronx' in borough_string.
7. Drop all rows that contain borough_num 0
8. Create a column 'processing_time_bucket' that contains 0 if processing time is less than 1 and 1 if it is greater or equal to 1.
9. Create a subset of the data that contains hour, month, weekday, agency_num, borough_num as independent variables and processing_time_bucket as the dependent variable
10. Save this subset of data as a CSV called 'filtered.csv'. The output should look something like that (don't forget the header):

```
hour,month,weekday,agency_num,borough_num,processing_t
ime_bucket
```

```
00,01,5,0,2,0

00,01,5,0,5,0

00,01,5,1,5,1

00,01,5,1,5,1

00,01,5,0,3,0

00,01,5,0,2,0

00,01,5,0,1,0
```

WARNING: Do not change the order of the rows. If you do, the grader won't recognize the data and you will get a low grade.

HINT: The file 'filtered.csv' should have 18792 rows, counting the header.

Question 2: prediction

Use a linear regression to classify the complaints in the 'topredict.csv' dataset. To do so, you should train a model on the data you saved in the 'filtered.csv' file. Save the predictions as a CSV called 'predictions.csv'.

You should have two columns, the row number ('index') of the complaint in the 'topredict.csv' file and the prediction ('prediction') you make for that complaint. The results should look like the following (don't forget the header):

```
index,prediction
```

```
0,0

1,0

2,1

3,0

4,0

5,1
```

What To Submit. Edit the starter code provided to solve this problem.

USE OF VOCAREUM

This assignment uses Vocareum for submission and grading. Vocareum comes equipped with an editing environment that you may use to do your development work. You are **NOT** required to use the editor. In particular, you are free to choose your favorite editor / IDE to do your development work on. When you are done with your work, you can simply upload your files onto Vocareum for submission and grading.

However, your assignments will be graded on the platform, so you **MUST** make sure that your code executes without error on the platform. To do so, you may use the **RUN** button to make sure that your code runs properly. In particular, we do not recommend using any third-party libraries and packages beyond what is imported in the starting code. We do not guarantee that they will work on the platform, even if they work on your personal computer. For the purposes of this project, the standard Python library and the imported packages should be sufficient.

Machine Learning Assignment - Vocareum (External resource)

(100.0 points possible)

Your email address will be used to identify your submission entry.

Launch Project

Learn About Verified Certificates