KDD2014

March 21, 2016

0.1 KDD2014 Project

import necessary library

```
In [1]: # imports
        import pandas as pd
        import matplotlib.pyplot as plt
        import csv
        import pandas as pd
        import numpy as np
        from sklearn.preprocessing import LabelEncoder, OneHotEncoder
        from sklearn.linear_model import LogisticRegression

        # this allows plots to appear directly in the notebook
        %matplotlib inline
```

load the data

```
In [2]: # load the data
        print('loading the data...')
        projects = pd.read_csv('./data/projects.csv')
        outcomes = pd.read_csv('./data/outcomes.csv')
        sample = pd.read_csv('./data/sampleSubmission.csv')
        print('complete..')
```

```
loading the data...
complete..
```

check data information

```
In [3]: print("projects", projects.shape)
        print("outcomes", outcomes.shape)
```

```
projects (664098, 35)
outcomes (619326, 12)
```

```
In [4]: projects.head(10)
```

```
Out[4]:                         projectid                     teacher_acctid  \
        0  316ed8fb3b81402ff6ac8f721bb31192  42d43fa6f37314365d08692e08680973
        1  90de744e368a7e4883223ca49318ae30  864eb466462bf704bf7a16a585ef296a
        2  32943bb1063267de6ed19fc0ceb4b9a7  37f85135259ece793213aca9d8765542
        3  bb18f409abda2f264d5acda8cab577a9  2133fc46f951f1e7d60645b0f9e48a6c
        4  24761b686e18e5eace634607acbcc19f  867ff478a63f5457eaf41049536c47cd
        5  eac7d156205f1333de3887d656f46611  ff064802c18e68db7ddb7ea0bf7732e8
```

```
6  5a3bfdf2e05781ccd0654dee0d51d1cd  085794a9e315b88cb7aec548831572f5
7  afda16eb54d8992db7bb42923b2a0c69  1d94a31c2dc38d51350eb26a4a2d892c
8  2ab3efb23acc84017cd7896f62c2e889  5a497c425e05bb193564ef3ce2cbf3b6
9  3118962680bb062323c3566197487315  eb0855cc6ea55d173ca8f8e4fd6a9e18

                           schoolid  school_ncesid  school_latitude  \
0  c0e6ce89b244764085691a1b8e28cb81   6.362701e+10        36.576340
1  d711e47810900c96f26a5d0be30c446d   4.837020e+11        32.911179
2  665c3613013ba0a66e3a2a26b89f1b68   4.103270e+11        45.166039
3  4f12c3fa0c1cce823c7ba1df57e90ccb   3.600153e+11        40.641727
4  10179fd362d7b8cf0e89baa1ca3025bb   6.227100e+10        34.043939
5  929336e5a242d2e67f1a591196ba7701   4.018700e+10        33.298792
6  dc34b021fd177cfdb22f22055ca9d04f   1.006000e+10        32.838778
7  86fff7caa07e4f595a716d8a1960f09e   1.301290e+11        33.896013
8  ed170a147cd8ee2391b54b2f8b206160   2.612000e+11        42.422054
9  71b4dee33799c2111773bb4444766bf0   3.402640e+11        39.951767

   school_longitude  school_city school_state  school_zip school_metro  \
0       -119.608713        Selma           CA       93662          NaN
1        -96.723640       Dallas           TX       75243        urban
2       -122.414576       Colton           OR       97017        rural
3        -73.965655     Brooklyn           NY       11226        urban
4       -118.288371  Los Angeles           CA       90006        urban
5       -111.827793     Chandler           AZ       85225     suburban
6        -85.517527    Lafayette           AL       36862        rural
7        -84.554213     Marietta           GA       30060          NaN
8        -83.209656      Detroit           MI       48235        urban
9        -75.117188       Camden           NJ        8102        urban

      ...     resource_type    poverty_level    grade_level  \
0     ...             Books  highest poverty     Grades 6-8
1     ...             Books  highest poverty  Grades PreK-2
2     ...        Technology     high poverty  Grades PreK-2
3     ...             Books  highest poverty     Grades 3-5
4     ...             Other  highest poverty  Grades PreK-2
5     ...        Technology  highest poverty     Grades 3-5
6     ...          Supplies  highest poverty  Grades PreK-2
7     ...             Books  highest poverty  Grades PreK-2
8     ...        Technology  highest poverty  Grades PreK-2
9     ...        Technology  highest poverty  Grades PreK-2

  fulfillment_labor_materials total_price_excluding_optional_support  \
0                          30                                 555.81
1                          30                                 296.47
2                          30                                 430.89
3                          30                                 576.07
4                          30                                 408.40
5                          30                                 750.92
6                          30                                2291.48
7                          30                                 459.36
8                          30                                 325.39
9                          30                                 567.84

  total_price_including_optional_support students_reached  \
0                                 653.89               32
1                                 348.79               22
2                                 506.93               17
3                                 677.73               12
4                                 480.47               24
5                                 883.44               20
6                                2695.86              320
7                                 540.42               18
8                                 382.81               25
9                                 668.05               28

  eligible_double_your_impact_match eligible_almost_home_match date_posted
0                                 f                          f  2014-05-12
1                                 f                          f  2014-05-12
2                                 f                          f  2014-05-11
3                                 f                          f  2014-05-11
4                                 f                          f  2014-05-11
5                                 f                          f  2014-05-11
6                                 f                          f  2014-05-11
7                                 f                          f  2014-05-11
8                                 f                          f  2014-05-11
9                                 f                          f  2014-05-11

[10 rows x 35 columns]
```

```
In [5]: outcomes.head()
```

```
Out[5]:
                          projectid is_exciting  \
0  ffffc4f85b60efc5b52347df489d0238           f
1  ffffac55ee02a49d1abc87ba6fc61135           f
2  ffff97ed93720407d70a2787475932b0           f
3  ffff418bb42fad24347527ad96100f81           f
4  ffff2d9c769c8fb5335e949c615425eb           t

  at_least_1_teacher_referred_donor fully_funded at_least_1_green_donation  \
0                               NaN            f                       NaN
1                                 f            t                         t
2                                 f            t                         t
3                                 f            f                         t
4                                 t            t                         t

  great_chat three_or_more_non_teacher_referred_donors  \
0          f                                       NaN
1          f                                         t
2          t                                         t
3          t                                         f
4          t                                         f

  one_non_teacher_referred_donor_giving_100_plus  \
0                                            NaN
1                                              f
2                                              t
3                                              f
4                                              t
```

```
  donation_from_thoughtful_donor  great_messages_proportion  \
0                            NaN                        NaN
1                              f                         57
2                              f                        100
3                              f                        100
4                              f                         63

   teacher_referred_count  non_teacher_referred_count
0                     NaN                         NaN
1                       0                           7
2                       0                           3
3                       0                           1
4                       6                           2
```

```
In [6]: # sort the data based on id
        projects = projects.sort('projectid')
        sample = sample.sort('projectid')
        outcomes = outcomes.sort('projectid')
```

```
/home/ubuntu/anaconda3/lib/python3.5/site-packages/ipykernel/__main__.py:2: FutureWarning: sort(columns=
  from ipykernel import kernelapp as app
/home/ubuntu/anaconda3/lib/python3.5/site-packages/ipykernel/__main__.py:3: FutureWarning: sort(columns=
  app.launch_new_instance()
/home/ubuntu/anaconda3/lib/python3.5/site-packages/ipykernel/__main__.py:4: FutureWarning: sort(columns=
```

```
In [7]: projects.head()
```

```
Out[7]:                                projectid                     teacher_acctid  \
        148979  00001ccc0e81598c4bd86bacb94d7acb  96963218e74e10c3764a5cfb153e6fea
        437277  00002bff514104264a6b798356fdd893  3414541eb63108700b188648f866f483
        405458  00002d691c05c51a5fdfbb2baef0ba25  7ad6abc974dd8b62773f79f6cbed48d5
        91352   0000b38bbc7252972f7984848cf58098  e1aa1ae5301d0cda860c4d9c89c24919
        49606   0000ee613c92ddc5298bf63142996a5c  e0c0a0214d3c2cfdc0ab6639bc3c5342

                                        schoolid  school_ncesid  school_latitude  \
        148979  9f3f9f2c2da7edda5648ccd10554ed8c   1.709930e+11        41.807654
        437277  cbaae3265eda78d330cb8ab1a9217071   6.032700e+10        35.203447
        405458  56502bae9e97bab5eb54f9001878f469   6.029700e+10        34.137997
        91352   30fcfca739b17be54ce3f1ee46980340   2.311400e+11        44.437717
        49606   38bb0d62aa613c2f933de56c9df855b7   5.101260e+11        38.851982

                school_longitude   school_city school_state  school_zip school_metro  \
        148979        -87.673257       Chicago           IL       60609        urban
        437277       -118.840956         Arvin           CA       93203          NaN
        405458       -118.062795       Arcadia           CA       91007        urban
        91352         -70.201292     Livermore           ME        4253        rural
        49606         -77.145287  Falls Church           VA       22041     suburban

                   ...     resource_type     poverty_level    grade_level  \
        148979     ...          Supplies   highest poverty  Grades PreK-2
        437277     ...          Supplies   highest poverty  Grades PreK-2
        405458     ...             Books  moderate poverty     Grades 3-5
        91352      ...        Technology   highest poverty  Grades PreK-2
        49606      ...        Technology      high poverty  Grades PreK-2
```

```
       fulfillment_labor_materials total_price_excluding_optional_support  \
148979                          30                                1273.82
437277                          35                                 477.32
405458                          35                                 892.31
91352                           30                                 547.86
49606                           30                                 384.86

       total_price_including_optional_support students_reached  \
148979                                1498.61               31
437277                                 561.55               20
405458                                1049.78              250
91352                                  644.54               36
49606                                  452.78               19

       eligible_double_your_impact_match eligible_almost_home_match  \
148979                                 f                          f
437277                                 t                          f
405458                                 f                          f
91352                                  t                          f
49606                                  f                          f

       date_posted
148979  2013-04-14
437277  2010-09-08
405458  2010-12-10
91352   2013-09-27
49606   2013-12-11

[5 rows x 35 columns]
```

check the missing data

```
In [8]: totalCount = projects.shape[0]

        for i in range(1,projects.shape[1]):
            nullcount = projects[projects[projects.columns[i]].isnull()].shape[0]
            percentage=float(nullcount)/float(totalCount) *100
            if(percentage>0):
                print(projects.columns[i],percentage,'%')
```

```
school_ncesid 6.4351948055859225 %
school_zip 0.0006023207418182256 %
school_metro 12.333721830211806 %
school_district 0.14275001581091948 %
school_county 0.002559863152727459 %
teacher_prefix 0.0006023207418182256 %
primary_focus_subject 0.0058726272327277 %
primary_focus_area 0.0058726272327277 %
secondary_focus_subject 31.304566494704094 %
secondary_focus_area 31.304566494704094 %
resource_type 0.006776108345455037 %
grade_level 0.0013552216690910076 %
fulfillment_labor_materials 5.282654066116748 %
students_reached 0.021984707076365236 %
```

fill up the missing data

```
In [9]: projects = projects.fillna(method='pad')
```

```
In [10]: projects.head(10)
```

```
Out[10]:                                projectid                    teacher_acctid  \
         148979  00001ccc0e81598c4bd86bacb94d7acb  96963218e74e10c3764a5cfb153e6fea
         437277  00002bff514104264a6b798356fdd893  3414541eb63108700b188648f866f483
         405458  00002d691c05c51a5fdfbb2baef0ba25  7ad6abc974dd8b62773f79f6cbed48d5
         91352   0000b38bbc7252972f7984848cf58098  e1aa1ae5301d0cda860c4d9c89c24919
         49606   0000ee613c92ddc5298bf63142996a5c  e0c0a0214d3c2cfdc0ab6639bc3c5342
         255442  0000fa3aa8f6649abab23615b546016d  2a578595fe351e7fce057e048c409b18
         189646  0000fb6aea57099cc5b051acb7f52a9e  ad51bb5eabffc738775887955421fe75
         616019  0001120447a33dd9ffeefa107ed04c43  a799e714a102967d674b258e5ea19231
         301504  0001146d343ea9452089d0e302496c06  3f71761d508f95684f2924763175dbe8
         511584  0001151477ea5349a0aa64ed1d83f0bc  f30b9edaea56bbade550e2f0da5db4f9

                                         schoolid  school_ncesid  school_latitude  \
         148979  9f3f9f2c2da7edda5648ccd10554ed8c   1.709930e+11        41.807654
         437277  cbaae3265eda78d330cb8ab1a9217071   6.032700e+10        35.203447
         405458  56502bae9e97bab5eb54f9001878f469   6.029700e+10        34.137997
         91352   30fcfca739b17be54ce3f1ee46980340   2.311400e+11        44.437717
         49606   38bb0d62aa613c2f933de56c9df855b7   5.101260e+11        38.851982
         255442  3432ed3d4466fac2f2ead83ab354e333   6.409801e+10        34.296596
         189646  d4f02777656b5ee806965ae2186e0adb   4.702940e+11        35.037663
         616019  c843a6322e90dc34304b60b43f4c2205   4.502580e+11        34.571828
         301504  18e8fc522b79044cf70938cbefce41bb   4.503870e+11        34.977550
         511584  6cd638cff9af07d02c72bb1cc25612d5   2.612000e+11        42.367172

                 school_longitude   school_city school_state  school_zip school_metro  \
         148979        -87.673257       Chicago           IL       60609        urban
         437277       -118.840956         Arvin           CA       93203        urban
         405458       -118.062795       Arcadia           CA       91007        urban
         91352         -70.201292     Livermore           ME        4253        rural
         49606         -77.145287  Falls Church           VA       22041     suburban
         255442       -119.296596       Ventura           CA       93001        urban
         189646        -90.092321       Memphis           TN       38109        urban
         616019        -80.615642       Kershaw           SC       29067        rural
         301504        -81.012395     Rock Hill           SC       29732        urban
         511584        -82.985527       Detroit           MI       48214        urban

                      ...     resource_type     poverty_level    grade_level  \
         148979       ...          Supplies   highest poverty  Grades PreK-2
         437277       ...          Supplies   highest poverty  Grades PreK-2
         405458       ...             Books  moderate poverty     Grades 3-5
         91352        ...        Technology   highest poverty  Grades PreK-2
         49606        ...        Technology      high poverty  Grades PreK-2
         255442       ...             Books   highest poverty     Grades 3-5
         189646       ...             Books   highest poverty     Grades 6-8
         616019       ...             Books      high poverty     Grades 6-8
         301504       ...        Technology      high poverty     Grades 3-5
         511584       ...          Supplies   highest poverty     Grades 3-5

                fulfillment_labor_materials total_price_excluding_optional_support  \
         148979                          30                                1273.82
         437277                          35                                 477.32
```

```
405458                          35                                 892.31
91352                           30                                 547.86
49606                           30                                 384.86
255442                          35                                 240.10
189646                          30                                 382.71
616019                          17                                 296.00
301504                          35                                 300.97
511584                           9                                 675.24

       total_price_including_optional_support students_reached  \
148979                                1498.61               31
437277                                 561.55               20
405458                                1049.78              250
91352                                  644.54               36
49606                                  452.78               19
255442                                 282.47               28
189646                                 450.25               90
616019                                 360.98               35
301504                                 354.08               21
511584                                 823.46              300

       eligible_double_your_impact_match eligible_almost_home_match  \
148979                                 f                          f
437277                                 t                          f
405458                                 f                          f
91352                                  t                          f
49606                                  f                          f
255442                                 t                          f
189646                                 f                          t
616019                                 f                          f
301504                                 f                          t
511584                                 f                          f

       date_posted
148979  2013-04-14
437277  2010-09-08
405458  2010-12-10
91352   2013-09-27
49606   2013-12-11
255442  2012-04-07
189646  2012-11-17
616019  2007-08-12
301504  2011-12-08
511584  2009-08-28

[10 rows x 35 columns]
```

```
In [11]: # split the training data and testing data
         dates = np.array(projects.date_posted)
```

```
In [12]: print(dates)
```

```
['2013-04-14' '2010-09-08' '2010-12-10' ..., '2010-09-11' '2011-06-11'
 '2009-10-11']
```

```
In [13]: train_idx = np.where(dates < '2014-01-01')[0]
         test_idx = np.where(dates >= '2014-01-01')[0]
```

```
In [14]: print(train_idx)
```

```
[     0      1      2 ..., 664095 664096 664097]
```

```
In [15]: print(test_idx)
```

```
[    33     52     53 ..., 664041 664088 664094]
```

```
In [16]: # check the data of training data
         print(projects.iloc[train_idx[9999]])
```

```
projectid                                 04253b89bf4b42f1e5f29b23418d36c3
teacher_acctid                            8aac3a46b9ac0df676cfc740d272a425
schoolid                                  8827b62394964f26a2764245dac80110
school_ncesid                                                   6.2808e+10
school_latitude                                                    37.9825
school_longitude                                                  -121.717
school_city                                                         Oakley
school_state                                                            CA
school_zip                                                           94561
school_metro                                                      suburban
school_district                                 Oakley Union Elem Sch Dist
school_county                                                 Contra Costa
school_charter                                                           f
school_magnet                                                            f
school_year_round                                                        t
school_nlns                                                              f
school_kipp                                                              f
school_charter_ready_promise                                            f
teacher_prefix                                                         Ms.
teacher_teach_for_america                                               f
teacher_ny_teaching_fellow                                              f
primary_focus_subject                                        Special Needs
primary_focus_area                                           Special Needs
secondary_focus_subject                                           Literacy
secondary_focus_area                                   Literacy & Language
resource_type                                                        Books
poverty_level                                             moderate poverty
grade_level                                                     Grades 3-5
fulfillment_labor_materials                                             30
total_price_excluding_optional_support                              199.08
total_price_including_optional_support                              234.21
students_reached                                                        10
eligible_double_your_impact_match                                       f
eligible_almost_home_match                                              f
date_posted                                                     2013-10-04
Name: 87071, dtype: object
```

```
In [17]: #preprocessing the data based on different types of attr
         projects_numeric_columns = ['school_latitude', 'school_longitude',
                                     'fulfillment_labor_materials',
                                     'total_price_excluding_optional_support',
                                     'total_price_including_optional_support']
```

```
In [18]: projects_id_columns = ['projectid', 'teacher_acctid', 'schoolid', 'school_ncesid']
```

```
In [19]: projects_categorial_columns = np.array(list(set(projects.columns).difference(set(projects_nume
```

```
In [20]: print(projects_categorial_columns)
```

```
['teacher_teach_for_america' 'school_charter_ready_promise' 'grade_level'
 'teacher_prefix' 'school_metro' 'poverty_level' 'primary_focus_area'
 'primary_focus_subject' 'school_kipp' 'resource_type' 'students_reached'
 'school_district' 'eligible_double_your_impact_match' 'school_city'
 'school_year_round' 'school_state' 'school_zip' 'school_nlns'
 'school_charter' 'eligible_almost_home_match' 'secondary_focus_area'
 'school_magnet' 'secondary_focus_subject' 'teacher_ny_teaching_fellow'
 'school_county']
```

```
In [21]: projects_categorial_values = np.array(projects[projects_categorial_columns])
```

```
In [22]: projects[projects_categorial_columns].head()
```

```
Out[22]:        teacher_teach_for_america school_charter_ready_promise    grade_level  \
         148979                         f                            f  Grades PreK-2
         437277                         f                            f  Grades PreK-2
         405458                         f                            f     Grades 3-5
         91352                          f                            f  Grades PreK-2
         49606                          f                            f  Grades PreK-2

                teacher_prefix school_metro     poverty_level   primary_focus_area  \
         148979           Mrs.        urban   highest poverty       Math & Science
         437277           Mrs.        urban   highest poverty  Literacy & Language
         405458            Mr.        urban  moderate poverty  Literacy & Language
         91352            Mrs.        rural   highest poverty  Literacy & Language
         49606             Ms.     suburban      high poverty  Literacy & Language

                primary_focus_subject school_kipp resource_type     ...      \
         148979           Mathematics           f      Supplies      ...
         437277              Literacy           f      Supplies      ...
         405458  Literature & Writing           f         Books      ...
         91352   Literature & Writing           f    Technology      ...
         49606                    ESL           f    Technology      ...

                 school_state school_zip school_nlns school_charter  \
         148979            IL      60609           f              f
         437277            CA      93203           f              f
         405458            CA      91007           f              f
         91352             ME       4253           f              f
         49606             VA      22041           f              f

                eligible_almost_home_match secondary_focus_area  school_magnet  \
         148979                          f     Music & The Arts              f
         437277                          f  Literacy & Language              f
         405458                          f  Literacy & Language              f
         91352                           f       Math & Science              f
         49606                           f        Special Needs              t

                secondary_focus_subject teacher_ny_teaching_fellow school_county
```

```
148979            Visual Arts                           f          Cook
437277    Literature & Writing                           f          Kern
405458                Literacy                           f   Los Angeles
91352              Mathematics                           f  Androscoggin
49606            Special Needs                           f       Fairfax

[5 rows x 25 columns]
```

```
In [23]: print(projects_categorial_values)
```

```
[[’f’ ’f’ ’Grades PreK-2’ ..., ’Visual Arts’ ’f’ ’Cook’]
 [’f’ ’f’ ’Grades PreK-2’ ..., ’Literature & Writing’ ’f’ ’Kern’]
 [’f’ ’f’ ’Grades 3-5’ ..., ’Literacy’ ’f’ ’Los Angeles’]
 ...,
 [’f’ ’f’ ’Grades PreK-2’ ..., ’Other’ ’f’ ’Kings (Brooklyn)’]
 [’f’ ’f’ ’Grades PreK-2’ ..., ’Literacy’ ’f’ ’Clayton’]
 [’f’ ’f’ ’Grades 3-5’ ..., ’Mathematics’ ’f’ ’Florence’]]
```

```
In [24]: # only use the category attr
         print(projects_categorial_columns)
         print(projects_categorial_columns.shape)
         print(projects_categorial_values[:, 0].shape)
```

```
[’teacher_teach_for_america’ ’school_charter_ready_promise’ ’grade_level’
 ’teacher_prefix’ ’school_metro’ ’poverty_level’ ’primary_focus_area’
 ’primary_focus_subject’ ’school_kipp’ ’resource_type’ ’students_reached’
 ’school_district’ ’eligible_double_your_impact_match’ ’school_city’
 ’school_year_round’ ’school_state’ ’school_zip’ ’school_nlns’
 ’school_charter’ ’eligible_almost_home_match’ ’secondary_focus_area’
 ’school_magnet’ ’secondary_focus_subject’ ’teacher_ny_teaching_fellow’
 ’school_county’]
(25,)
(664098,)
```

```
In [25]: # encode the category value and reform the original data
         label_encoder = LabelEncoder()

         # set up the encoding model, using the first row of data
         projects_data = label_encoder.fit_transform(projects_categorial_values[:,0])
```

```
In [26]: # use the model to transform the following data
         for i in range(1, projects_categorial_values.shape[1]):
             label_encoder = LabelEncoder()
             projects_data = np.column_stack((projects_data, label_encoder.fit_transform(projects_catego
```

```
In [27]: projects_data = projects_data.astype(float)
         print(’The shape of the project data’, projects_data.shape)
```

```
The shape of the project data (664098, 25)
```

```
In [28]: # one hot encoding
         enc = OneHotEncoder()
         enc.fit(projects_data)
         projects_data = enc.transform(projects_data)
```

```
In [29]: print(projects_data)
```

```
(0, 36457)        1.0
  (0, 36090)        1.0
  (0, 36089)        1.0
  (0, 36061)        1.0
  (0, 36059)        1.0
  (0, 36052)        1.0
  (0, 36050)        1.0
  (0, 36048)        1.0
  (0, 29542)        1.0
  (0, 19391)        1.0
  (0, 19375)        1.0
  (0, 11779)        1.0
  (0, 10366)        1.0
  (0, 7439)        1.0
  (0, 93)        1.0
  (0, 58)        1.0
  (0, 54)        1.0
  (0, 44)        1.0
  (0, 24)        1.0
  (0, 17)        1.0
  (0, 15)        1.0
  (0, 11)        1.0
  (0, 7)        1.0
  (0, 2)        1.0
  (0, 0)        1.0
  :        :
  (664097, 36627)        1.0
  (664097, 36090)        1.0
  (664097, 36080)        1.0
  (664097, 36061)        1.0
  (664097, 36058)        1.0
  (664097, 36052)        1.0
  (664097, 36050)        1.0
  (664097, 36048)        1.0
  (664097, 23853)        1.0
  (664097, 19418)        1.0
  (664097, 19375)        1.0
  (664097, 13051)        1.0
  (664097, 10366)        1.0
  (664097, 3859)        1.0
  (664097, 152)        1.0
  (664097, 59)        1.0
  (664097, 54)        1.0
  (664097, 44)        1.0
  (664097, 24)        1.0
  (664097, 16)        1.0
  (664097, 15)        1.0
  (664097, 11)        1.0
  (664097, 4)        1.0
  (664097, 2)        1.0
  (664097, 0)        1.0
```

```
In [30]: #Predicting
         train = projects_data[train_idx]
         test = projects_data[test_idx]
         print('shape of test', test.shape)
         clf = LogisticRegression()

shape of test (44772, 37794)

In [31]: print(test)

(0, 0)        1.0
  (0, 2)        1.0
  (0, 6)        1.0
  (0, 9)        1.0
  (0, 13)        1.0
  (0, 16)        1.0
  (0, 20)        1.0
  (0, 30)        1.0
  (0, 54)        1.0
  (0, 59)        1.0
  (0, 262)        1.0
  (0, 8171)        1.0
  (0, 10366)        1.0
  (0, 12981)        1.0
  (0, 19375)        1.0
  (0, 19399)        1.0
  (0, 20111)        1.0
  (0, 36048)        1.0
  (0, 36050)        1.0
  (0, 36053)        1.0
  (0, 36056)        1.0
  (0, 36061)        1.0
  (0, 36086)        1.0
  (0, 36090)        1.0
  (0, 36638)        1.0
  :        :
  (44771, 0)        1.0
  (44771, 2)        1.0
  (44771, 7)        1.0
  (44771, 11)        1.0
  (44771, 13)        1.0
  (44771, 16)        1.0
  (44771, 23)        1.0
  (44771, 42)        1.0
  (44771, 54)        1.0
  (44771, 57)        1.0
  (44771, 86)        1.0
  (44771, 7786)        1.0
  (44771, 10366)        1.0
  (44771, 17894)        1.0
  (44771, 19375)        1.0
  (44771, 19405)        1.0
  (44771, 23207)        1.0
  (44771, 36048)        1.0
  (44771, 36050)        1.0
  (44771, 36052)        1.0
  (44771, 36057)        1.0
  (44771, 36061)        1.0
  (44771, 36079)        1.0
  (44771, 36090)        1.0
  (44771, 37348)        1.0
```

```
In [32]: # set the target labels
         labels = np.array(outcomes.is_exciting)
         clf.fit(train, labels=='t')
```

```
Out[32]: LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
                   intercept_scaling=1, max_iter=100, multi_class='ovr', n_jobs=1,
                   penalty='l2', random_state=None, solver='liblinear', tol=0.0001,
                   verbose=0, warm_start=False)
```

```
In [33]: # perform the prediction
         preds = clf.predict(test)
```

```
In [34]: print(preds.shape)
```

```
(44772,)
```

```
In [35]: print(preds)
```

```
[False False False ..., False False False]
```

```
In [36]: f = lambda x : 1 if x else 0
         f = np.vectorize(f)
         preds = f(preds)
         preds.astype(int)
         print(preds)
```

```
[0 0 0 ..., 0 0 0]
```

```
In [37]: #Save prediction into a file
         sample['is_exciting'] = preds
```

```
In [38]: sample.head()
```

```
Out[38]:                               projectid  is_exciting
         44771  00034e54ed99042609edad55031c8861            0
         44770  00057f424b5498c7ece13d13ce3e2178            0
         44769  00059e63bb0567708b2b0c9e3d9c43d6            0
         44768  0008c67f27dd29ea7be5a7cc5a866df8            0
         44767  000b6e707ad50a597ab46eedff6bde05            0
```

```
In [39]: sample.tail()
```

```
Out[39]:                           projectid  is_exciting
         4  fff745e9c0b8cc9e73e8c4c9a0ef4292            0
         3  fff8beec6de8c9411520d15d1f6979bf            0
         2  fff979abefa35a6bdd133b4e4150b737            0
         1  fffeb510ee37a0bb01079f06bf141246            0
         0  ffff7266778f71242675416e600b94e1            0
```

```
In [40]: sample.to_csv('predictions.csv', index = False)
```

```
In [ ]:
```