In [87]:
```python
import pandas as pd
import numpy as np
import pandas as pd
import os

import matplotlib.pyplot as plt
import seaborn as sns

from sklearn.preprocessing import MinMaxScaler
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import plot_confusion_matrix
from sklearn.metrics import confusion_matrix, classification_report
from sklearn.metrics import roc_curve, auc
from sklearn.metrics import precision_score, recall_score, accuracy_score, f1_sco
from sklearn.preprocessing import StandardScaler
from sklearn.neighbors import KNeighborsClassifier
from sklearn.tree import DecisionTreeClassifier
from sklearn.tree import plot_tree
from sklearn.utils import resample
from imblearn.over_sampling import SMOTE

from sklearn.ensemble import BaggingClassifier, RandomForestClassifier
from sklearn.model_selection import GridSearchCV, cross_val_score
from sklearn.ensemble import AdaBoostClassifier, GradientBoostingClassifier
from xgboost import XGBClassifier

from sklearn import tree
import xgboost as xgb
from numpy import loadtxt
from xgboost import XGBClassifier
from xgboost import plot_tree

import gc
from tqdm import tqdm

from sklearn.preprocessing import StandardScaler, OneHotEncoder
from sklearn.impute import SimpleImputer
from sklearn.metrics import mean_squared_error
from sklearn.model_selection import cross_val_predict
from keras import models
from keras import layers
from keras import regularizers
from keras.wrappers.scikit_learn import KerasRegressor
```

executed in 32ms, finished 03:11:42 2021-04-09

In [88]:
```python
column_defs = pd.read_excel('data\LCDataDictionary.xlsx',index_col='LoanStatNew')
column_defs.columns
```

executed in 59ms, finished 03:11:43 2021-04-09

Out[88]:
```
Index(['Description'], dtype='object')
```

In [89]:
```
def na_check(data):
    check = np.round(data.isna().sum().sort_values(ascending=False),2)
    return check
```
executed in 15ms, finished 03:11:44 2021-04-09

In [90]:
```
def column_info(col_name):
    return column_defs.loc[col_name]['Description']
```
executed in 15ms, finished 03:11:45 2021-04-09

In [91]:
```python
def reduce_mem_usage(df, int_cast=True, obj_to_category=False, subset=None):
    """
    Iterate through all the columns of a dataframe and modify the data type to re
    :param df: dataframe to reduce (pd.DataFrame)
    :param int_cast: indicate if columns should be tried to be casted to int (boo
    :param obj_to_category: convert non-datetime related objects to category dtyp
    :param subset: subset of columns to analyse (list)
    :return: dataset with the column dtypes adjusted (pd.DataFrame)
    """
    start_mem = df.memory_usage().sum() / 1024 ** 2;
    gc.collect()
    print('Memory usage of dataframe is {:.2f} MB'.format(start_mem))

    cols = subset if subset is not None else df.columns.tolist()

    for col in tqdm(cols):
        col_type = df[col].dtype

        if col_type != object and col_type.name != 'category' and 'datetime' not
            c_min = df[col].min()
            c_max = df[col].max()

            # test if column can be converted to an integer
            treat_as_int = str(col_type)[:3] == 'int'
            if int_cast and not treat_as_int:
                treat_as_int = check_if_integer(df[col])

            if treat_as_int:
                if c_min > np.iinfo(np.int8).min and c_max < np.iinfo(np.int8).ma
                    df[col] = df[col].astype(np.int8)
                elif c_min > np.iinfo(np.uint8).min and c_max < np.iinfo(np.uint8
                    df[col] = df[col].astype(np.uint8)
                elif c_min > np.iinfo(np.int16).min and c_max < np.iinfo(np.int16
                    df[col] = df[col].astype(np.int16)
                elif c_min > np.iinfo(np.uint16).min and c_max < np.iinfo(np.uint
                    df[col] = df[col].astype(np.uint16)
                elif c_min > np.iinfo(np.int32).min and c_max < np.iinfo(np.int32
                    df[col] = df[col].astype(np.int32)
                elif c_min > np.iinfo(np.uint32).min and c_max < np.iinfo(np.uint
                    df[col] = df[col].astype(np.uint32)
                elif c_min > np.iinfo(np.int64).min and c_max < np.iinfo(np.int64
                    df[col] = df[col].astype(np.int64)
                elif c_min > np.iinfo(np.uint64).min and c_max < np.iinfo(np.uint
                    df[col] = df[col].astype(np.uint64)
            else:
                if c_min > np.finfo(np.float32).min and c_max < np.finfo(np.float
                    df[col] = df[col].astype(np.float32)
                else:
                    df[col] = df[col].astype(np.float64)
        elif 'datetime' not in col_type.name and obj_to_category:
            df[col] = df[col].astype('category')
    gc.collect()
    end_mem = df.memory_usage().sum() / 1024 ** 2
    print('Memory usage after optimization is: {:.3f} MB'.format(end_mem))
    print('Decreased by {:.1f}%'.format(100 * (start_mem - end_mem) / start_mem))
```

```python
    return df
```

executed in 27ms, finished 03:11:45 2021-04-09

In [113]:
```python
df = pd.read_csv('data/preprocessed.csv')
```

executed in 15.0s, finished 11:11:34 2021-04-12

In [114]:
```python
reduce_mem_usage(df,int_cast=False)
```

executed in 11.8s, finished 11:11:47 2021-04-12

```
  0%|          | 0/75 [00:00<?, ?it/s]

Memory usage of dataframe is 993.88 MB

100%|██████████| 75/75 [00:09<00:00,  7.56it/s]

Memory usage after optimization is: 631.116 MB
Decreased by 36.5%
```

Out[114]:

	Unnamed: 0	Unnamed: 0.1	id	loan_amnt	term	int_rate	installment	grac
0	0	42536	10129454	12000.0	36 months	10.99%	392.799988	
1	1	42537	10149488	4800.0	36 months	10.99%	157.100006	
2	2	42538	10149342	27060.0	36 months	10.99%	885.500000	
3	3	42539	10148122	12000.0	36 months	7.62%	374.000000	
4	4	42540	10129477	14000.0	36 months	12.85%	470.799988	
...	...	...	...	...	...	...	...	
1736932	1741058	2925488	102556443	24000.0	60 months	23.99%	690.500000	
1736933	1741059	2925489	102653304	10000.0	36 months	7.99%	313.200012	
1736934	1741060	2925490	102628603	10050.0	36 months	16.99%	358.200012	
1736935	1741061	2925491	102196576	6000.0	36 months	11.44%	197.800003	
1736936	1741062	2925492	99799684	30000.0	60 months	25.49%	889.000000	

1736937 rows × 75 columns

In [18]:
```
for col in df.columns:
    print(col)
    print(column_info(col))
    print(df[col].value_counts(normalize = True, ascending=False).head(3))
    print("--------------------------------------------------------------")
```
executed in 2.78s, finished 17:11:58 2021-04-05

```
id
A unique LC assigned ID for the loan listing.
4196351      5.757261e-07
75101579     5.757261e-07
137966995    5.757261e-07
Name: id, dtype: float64
--------------------------------------------------------------
loan_amnt
The listed amount of the loan applied for by the borrower. If at some point i
n time, the credit department reduces the loan amount, then it will be reflec
ted in this value.
10000.0    0.078220
20000.0    0.054624
12000.0    0.053229
Name: loan_amnt, dtype: float64
--------------------------------------------------------------
term
The number of payments on the loan. Values are in months and can be either 36
or 60.
```

In [94]:
```
df.drop(columns=['Unnamed: 0','Unnamed: 0.1'],axis=1,inplace=True)
```
executed in 655ms, finished 03:12:17 2021-04-09

Have cleaned for NA values but after baseline model we need to drop more variables as our model is overfitting

will look to drop features that are not relevant to performance of the loans as they are not available at time of application approval

Payment plan and out_principal only have 1 value so dropping

In [95]:
```
df.drop(columns=['pymnt_plan','out_prncp'],axis=1,inplace=True)
```
executed in 422ms, finished 03:12:21 2021-04-09

In [96]:
```
post_app_drops = ['total_bal_ex_mort', 'pct_tl_nvr_dlq','num_tl_op_past_12m','nur
,'num_tl_120dpd_2m','num_bc_sats','num_accts_ever_120_pd','mths_since_recent_bc',
,'chargeoff_within_12_mths','bc_util','avg_cur_bal','tot_cur_bal','tot_coll_amt',
,'last_fico_range_low','last_fico_range_high','last_credit_pull_d','last_pymnt_ar
,'total_rec_int','total_rec_prncp','total_pymnt','revol_util','revol_bal']
```
executed in 15ms, finished 03:12:21 2021-04-09

In [15]:
```
for x in post_app_drops:
    print('{}: {}'.format(x,column_info(x)))
```
executed in 25ms, finished 18:16:20 2021-04-07

```
total_bal_ex_mort: Total credit balance excluding mortgage
pct_tl_nvr_dlq: Percent of trades never delinquent
num_tl_op_past_12m: Number of accounts opened in past 12 months
num_tl_90g_dpd_24m: Number of accounts 90 or more days past due in last 24 mont
hs
num_tl_30dpd: Number of accounts currently 30 days past due (updated in past 2
months)
num_tl_120dpd_2m: Number of accounts currently 120 days past due (updated in pa
st 2 months)
num_bc_sats: Number of satisfactory bankcard accounts
num_accts_ever_120_pd: Number of accounts ever 120 or more days past due
mths_since_recent_bc: Months since most recent bankcard account opened.
mo_sin_rcnt_tl: Months since most recent account opened
mo_sin_rcnt_rev_tl_op: Months since most recent revolving account opened
mo_sin_old_rev_tl_op: Months since oldest revolving account opened
chargeoff_within_12_mths: Number of charge-offs within 12 months
bc_util: Ratio of total current balance to high credit/credit limit for all ban
kcard accounts.
avg_cur_bal: Average current balance of all accounts
tot_cur_bal: Total current balance of all accounts
tot_coll_amt: Total collection amounts ever owed
acc_now_delinq: The number of accounts on which the borrower is now delinquent.
collections_12_mths_ex_med: Number of collections in 12 months excluding medica
l collections
last_fico_range_low: The lower boundary range the borrower’s last FICO pulled b
elongs to.
last_fico_range_high: The upper boundary range the borrower’s last FICO pulled
belongs to.
last_credit_pull_d: The most recent month LC pulled credit for this loan
last_pymnt_amnt: Last total payment amount received
last_pymnt_d: Last month payment was received
recoveries: post charge off gross recovery
total_rec_late_fee: Late fees received to date
total_rec_int: Interest received to date
total_rec_prncp: Principal received to date
total_pymnt: Payments received to date for total amount funded
revol_util: Revolving line utilization rate, or the amount of credit the borrow
er is using relative to all available revolving credit.
revol_bal: Total credit revolving balance
```

In [97]:
```
df.drop(columns=post_app_drops,axis=1,inplace=True)
```
executed in 324ms, finished 03:12:27 2021-04-09

In [17]:
```
df.info()
```
executed in 13ms, finished 18:17:56 2021-04-07

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1736937 entries, 0 to 1736936
Data columns (total 40 columns):
 #   Column                       Dtype  
---  ------                       -----  
 0   id                           int32  
 1   loan_amnt                    float32
 2   term                         object 
 3   int_rate                     object 
 4   installment                  float32
 5   grade                        object 
 6   sub_grade                    object 
 7   emp_length                   int8   
 8   home_ownership               object 
 9   annual_inc                   float32
 10  verification_status          object 
 11  issue_d                      object 
 12  loan_status                  object 
 13  purpose                      object 
 14  title                        object 
 15  zip_code                     object 
 16  addr_state                   object 
 17  dti                          float32
 18  delinq_2yrs                  float32
 19  earliest_cr_line             object 
 20  fico_range_low               float32
 21  fico_range_high              float32
 22  inq_last_6mths               float32
 23  pub_rec                      float32
 24  total_acc                    float32
 25  initial_list_status          object 
 26  application_type             object 
 27  acc_open_past_24mths         float32
 28  delinq_amnt                  float32
 29  mort_acc                     float32
 30  mths_since_recent_inq        float32
 31  num_bc_tl                    float32
 32  num_il_tl                    float32
 33  num_op_rev_tl                float32
 34  pub_rec_bankruptcies         float32
 35  tax_liens                    float32
 36  total_bc_limit               float32
 37  total_il_high_credit_limit   float32
 38  hardship_flag                object 
 39  debt_settlement_flag         object 
dtypes: float32(21), int32(1), int8(1), object(17)
memory usage: 372.7+ MB
```

In [13]:
```python
for col in df.columns:
    print(col)
    print(column_info(col))
    print(df[col].value_counts(normalize = True, ascending=False).head(3))
    print("-------------------------------------------------------------")
```

executed in 2.88s, finished 17:38:24 2021-04-05

```
id
A unique LC assigned ID for the loan listing.
4196351      5.757261e-07
75101579     5.757261e-07
137966995    5.757261e-07
Name: id, dtype: float64
-------------------------------------------------------------
loan_amnt
The listed amount of the loan applied for by the borrower. If at some point in
time, the credit department reduces the loan amount, then it will be reflected
in this value.
10000.0    0.078220
20000.0    0.054624
12000.0    0.053229
Name: loan_amnt, dtype: float64
-------------------------------------------------------------
term
The number of payments on the loan. Values are in months and can be either 36 o
r 60.
 36 months    0.748299
 60 months    0.251701
Name: term, dtype: float64
-------------------------------------------------------------
int_rate
Interest Rate on the loan
  5.32%    0.023535
 10.99%    0.023520
 11.99%    0.023327
Name: int_rate, dtype: float64
-------------------------------------------------------------
installment
The monthly payment owed by the borrower if the loan originates.
361.500000    0.002526
301.200012    0.002434
318.799988    0.002335
Name: installment, dtype: float64
-------------------------------------------------------------
grade
LC assigned loan grade
B    0.293626
C    0.286439
A    0.185455
Name: grade, dtype: float64
-------------------------------------------------------------
sub_grade
LC assigned loan subgrade
C1    0.064049
B5    0.062420
B4    0.062224
Name: sub_grade, dtype: float64
```

```
-------------------------------------------------------------
emp_length
Employment length in years. Possible values are between 0 and 10 where 0 means
less than one year and 10 means ten or more years.
10    0.353702
1     0.157524
2     0.096483
Name: emp_length, dtype: float64
-------------------------------------------------------------
home_ownership
The home ownership status provided by the borrower during registration or obtai
ned from the credit report. Our values are: RENT, OWN, MORTGAGE, OTHER
MORTGAGE    0.495695
RENT        0.392156
OWN         0.111446
Name: home_ownership, dtype: float64
-------------------------------------------------------------
annual_inc
The self-reported annual income provided by the borrower during registration.
60000.0    0.038653
50000.0    0.033990
65000.0    0.029173
Name: annual_inc, dtype: float64
-------------------------------------------------------------
verification_status
Indicates if income was verified by LC, not verified, or if the income source w
as verified
Source Verified    0.399591
Not Verified       0.314003
Verified           0.286405
Name: verification_status, dtype: float64
-------------------------------------------------------------
issue_d
The month which the loan was funded
2016-03-01    0.030292
2015-10-01    0.025837
2015-07-01    0.024701
Name: issue_d, dtype: float64
-------------------------------------------------------------
loan_status
Current status of the loan
Fully Paid     0.805926
Charged Off    0.194074
Name: loan_status, dtype: float64
-------------------------------------------------------------
purpose
A category provided by the borrower for the loan request.
debt_consolidation    0.574522
credit_card           0.226116
home_improvement      0.066082
Name: purpose, dtype: float64
-------------------------------------------------------------
title
The loan title provided by the borrower
Debt consolidation         0.530573
Credit card refinancing    0.208071
Home improvement           0.062103
```

```
Name: title, dtype: float64
-------------------------------------------------------------
zip_code
The first 3 numbers of the zip code provided by the borrower in the loan applic
ation.
945xx    0.010728
750xx    0.010697
112xx    0.010148
Name: zip_code, dtype: float64
-------------------------------------------------------------
addr_state
The state provided by the borrower in the loan application
CA    0.141072
TX    0.082699
NY    0.080518
Name: addr_state, dtype: float64
-------------------------------------------------------------
dti
A ratio calculated using the borrower’s total monthly debt payments on the tota
l debt obligations, excluding mortgage and the requested LC loan, divided by th
e borrower’s self-reported monthly income.
16.799999    0.000993
19.200001    0.000946
17.700001    0.000917
Name: dti, dtype: float64
-------------------------------------------------------------
delinq_2yrs
The number of 30+ days past-due incidences of delinquency in the borrower's cre
dit file for the past 2 years
0.0    0.808393
1.0    0.127379
2.0    0.037085
Name: delinq_2yrs, dtype: float64
-------------------------------------------------------------
earliest_cr_line
The month the borrower's earliest reported credit line was opened
Sep-2004    0.006813
Sep-2003    0.006806
Aug-2001    0.006619
Name: earliest_cr_line, dtype: float64
-------------------------------------------------------------
fico_range_low
The lower boundary range the borrower’s FICO at loan origination belongs to.
660.0    0.087409
670.0    0.084559
665.0    0.084192
Name: fico_range_low, dtype: float64
-------------------------------------------------------------
fico_range_high
The upper boundary range the borrower’s FICO at loan origination belongs to.
664.0    0.087409
674.0    0.084559
669.0    0.084192
Name: fico_range_high, dtype: float64
-------------------------------------------------------------
inq_last_6mths
The number of inquiries in past 6 months (excluding auto and mortgage inquirie
```

```
s)
0.0    0.594151
1.0    0.267264
2.0    0.093499
Name: inq_last_6mths, dtype: float64
-------------------------------------------------------------
pub_rec
Number of derogatory public records
0.0    0.830046
1.0    0.143627
2.0    0.017168
Name: pub_rec, dtype: float64
-------------------------------------------------------------
total_acc
The total number of credit lines currently in the borrower's credit file
20.0    0.036317
19.0    0.036039
21.0    0.035920
Name: total_acc, dtype: float64
-------------------------------------------------------------
initial_list_status
The initial listing status of the loan. Possible values are – W, F
w    0.664308
f    0.335692
Name: initial_list_status, dtype: float64
-------------------------------------------------------------
application_type
Indicates whether the loan is an individual application or a joint application
with two co-borrowers
Individual    0.962883
Joint App     0.037117
Name: application_type, dtype: float64
-------------------------------------------------------------
acc_open_past_24mths
Number of trades opened in past 24 months.
3.0    0.147369
4.0    0.139893
2.0    0.132013
Name: acc_open_past_24mths, dtype: float64
-------------------------------------------------------------
delinq_amnt
The past-due amount owed for the accounts on which the borrower is now delinque
nt.
0.0        0.996464
25.0       0.000060
65000.0    0.000056
Name: delinq_amnt, dtype: float64
-------------------------------------------------------------
mort_acc
Number of mortgage accounts.
0.0    0.411127
1.0    0.176516
2.0    0.147438
Name: mort_acc, dtype: float64
-------------------------------------------------------------
mths_since_recent_inq
Months since most recent inquiry.
```

```
1.0    0.111902
2.0    0.090702
0.0    0.088569
Name: mths_since_recent_inq, dtype: float64
-------------------------------------------------------------
num_bc_tl
Number of bankcard accounts
5.0    0.103844
6.0    0.101556
4.0    0.098095
Name: num_bc_tl, dtype: float64
-------------------------------------------------------------
num_il_tl
Number of installment accounts
4.0    0.085617
3.0    0.085102
5.0    0.081857
Name: num_il_tl, dtype: float64
-------------------------------------------------------------
num_op_rev_tl
Number of open revolving accounts
6.0    0.109214
5.0    0.105886
7.0    0.103536
Name: num_op_rev_tl, dtype: float64
-------------------------------------------------------------
pub_rec_bankruptcies
Number of public record bankruptcies
0.0    0.872621
1.0    0.120449
2.0    0.005413
Name: pub_rec_bankruptcies, dtype: float64
-------------------------------------------------------------
tax_liens
Number of tax liens
0.0    0.968446
1.0    0.021345
2.0    0.005909
Name: tax_liens, dtype: float64
-------------------------------------------------------------
total_bc_limit
Total bankcard high credit/credit limit
5000.0     0.007954
6000.0     0.007106
10000.0    0.006792
Name: total_bc_limit, dtype: float64
-------------------------------------------------------------
total_il_high_credit_limit
Total installment high credit/credit limit
0.0        0.119217
10000.0    0.005989
15000.0    0.004659
Name: total_il_high_credit_limit, dtype: float64
-------------------------------------------------------------
hardship_flag
Flags whether or not the borrower is on a hardship plan
```

```
N    0.999073
Y    0.000927
Name: hardship_flag, dtype: float64
-------------------------------------------------------------
debt_settlement_flag
Flags whether or not the borrower, who has charged-off, is working with a debt-
settlement company.
N    0.971706
Y    0.028294
Name: debt_settlement_flag, dtype: float64
-------------------------------------------------------------
```

In [30]:
```
df.sub_grade
```
executed in 40ms, finished 13:10:48 2021-04-08

Out[30]:
```
0          B2
1          B2
2          B2
3          A3
4          B4
           ..
1736932    E2
1736933    A5
1736934    D1
1736935    B4
1736936    E4
Name: sub_grade, Length: 1736937, dtype: object
```

In [98]:
```
df.drop(columns=["debt_settlement_flag", "hardship_flag",'id'],axis=1,inplace=Tru
```
executed in 275ms, finished 03:12:35 2021-04-09

In [99]:
```
# making average fico score and dropping the fico range high and low
df['average_fico'] = (df['fico_range_high'] + df['fico_range_low'])/2
df.drop(columns=['fico_range_high','fico_range_low'],axis=1,inplace=True)
```
executed in 411ms, finished 03:12:36 2021-04-09

In [16]:
```
df.info()
```
executed in 8ms, finished 17:38:45 2021-04-05

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1736937 entries, 0 to 1736936
Data columns (total 36 columns):
 #   Column                       Dtype  
---  ------                       -----  
 0   loan_amnt                    float32
 1   term                         object 
 2   int_rate                     object 
 3   installment                  float32
 4   grade                        object 
 5   sub_grade                    object 
 6   emp_length                   int8   
 7   home_ownership               object 
 8   annual_inc                   float32
 9   verification_status          object 
 10  issue_d                      object 
 11  loan_status                  object 
 12  purpose                      object 
 13  title                        object 
 14  zip_code                     object 
 15  addr_state                   object 
 16  dti                          float32
 17  delinq_2yrs                  float32
 18  earliest_cr_line             object 
 19  inq_last_6mths               float32
 20  pub_rec                      float32
 21  total_acc                    float32
 22  initial_list_status          object 
 23  application_type             object 
 24  acc_open_past_24mths         float32
 25  delinq_amnt                  float32
 26  mort_acc                     float32
 27  mths_since_recent_inq        float32
 28  num_bc_tl                    float32
 29  num_il_tl                    float32
 30  num_op_rev_tl                float32
 31  pub_rec_bankruptcies         float32
 32  tax_liens                    float32
 33  total_bc_limit               float32
 34  total_il_high_credit_limit   float32
 35  average_fico                 float32
dtypes: float32(20), int8(1), object(15)
memory usage: 333.0+ MB
```

Export to csv from here rest is seeing if improvement to baseline

In [100]:
```
df.drop('initial_list_status',axis=1,inplace=True)
```
executed in 252ms, finished 03:12:39 2021-04-09

In [101]:
```
df.drop('title',axis=1,inplace=True)
```
executed in 256ms, finished 03:12:40 2021-04-09

In [102]:
```
df.drop(columns=['zip_code'],axis=1,inplace=True)
```
executed in 231ms, finished 03:12:45 2021-04-09

In [103]:
```
df.int_rate = df.int_rate.map(lambda x: float(x.replace('%','')))
```
executed in 845ms, finished 03:12:47 2021-04-09

In [28]:
```
df.loan_status.value_counts()
```
executed in 124ms, finished 18:20:46 2021-04-07

Out[28]:
```
Fully Paid     1399842
Charged Off     337095
Name: loan_status, dtype: int64
```

1 Addr State

Here converting states to regions as using state dummy variables creates 50 variables

In [104]:
```
regions = pd.read_excel('data/state_regions.xlsx')
```
executed in 26ms, finished 03:13:05 2021-04-09

In [105]:
```
df.addr_state
```
executed in 8ms, finished 03:13:12 2021-04-09

Out[105]:
```
0          NC
1          TX
2          MI
3          TX
4          NC
           ..
1736932    CO
1736933    PA
1736934    VA
1736935    NY
1736936    TX
Name: addr_state, Length: 1736937, dtype: object
```

In [106]:
```
df.addr_state.iloc[0]
```
executed in 7ms, finished 03:13:41 2021-04-09

Out[106]:
```
'NC'
```

In [107]:
```
regions[regions['State Code'] == 'AK']['Region']
```
executed in 14ms, finished 03:13:42 2021-04-09

Out[107]:
```
0    West
Name: Region, dtype: object
```

In [108]:
```
df['region'] = df.addr_state.apply(lambda x: regions.loc[regions['State Code'] ==
```
executed in 7m 25s, finished 03:21:10 2021-04-09

In [109]:
```
df.region.value_counts()
```
executed in 109ms, finished 03:22:40 2021-04-09

Out[109]:
```
South        621948
West         456308
Northeast    349635
Midwest      309046
Name: region, dtype: int64
```

In [111]:
```
df.drop(columns = ['addr_state'],axis=1,inplace=True)
```
executed in 394ms, finished 03:45:20 2021-04-09

In [110]:
```
df.info()
```
executed in 20ms, finished 03:22:45 2021-04-09

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1736937 entries, 0 to 1736936
Data columns (total 34 columns):
 #   Column                      Dtype  
---  ------                      -----  
 0   loan_amnt                   float32
 1   term                        object 
 2   int_rate                    float64
 3   installment                 float32
 4   grade                       object 
 5   sub_grade                   object 
 6   emp_length                  int8   
 7   home_ownership              object 
 8   annual_inc                  float32
 9   verification_status         object 
 10  issue_d                     object 
 11  loan_status                 object 
 12  purpose                     object 
 13  addr_state                  object 
 14  dti                         float32
 15  delinq_2yrs                 float32
 16  earliest_cr_line            object 
 17  inq_last_6mths              float32
 18  pub_rec                     float32
 19  total_acc                   float32
 20  application_type            object 
 21  acc_open_past_24mths        float32
 22  delinq_amnt                 float32
 23  mort_acc                    float32
 24  mths_since_recent_inq       float32
 25  num_bc_tl                   float32
 26  num_il_tl                   float32
 27  num_op_rev_tl               float32
 28  pub_rec_bankruptcies        float32
 29  tax_liens                   float32
 30  total_bc_limit              float32
 31  total_il_high_credit_limit  float32
 32  average_fico                float32
 33  region                      object 
dtypes: float32(20), float64(1), int8(1), object(12)
memory usage: 306.4+ MB
```

In [ ]:

In [ ]:

In [ ]:

In [112]:
```
df.to_csv('data/full_clean')
```
executed in 31.0s, finished 03:45:55 2021-04-09

2 Baseline Test

In [83]:
```
df.drop(columns = ['regions','addr_state'],axis=1,inplace=True)
```
executed in 378ms, finished 02:16:42 2021-04-09

In [20]:
```
df.columns
```
executed in 9ms, finished 17:43:01 2021-04-05

Out[20]:
```
Index(['loan_amnt', 'term', 'int_rate', 'installment', 'sub_grade',
       'emp_length', 'home_ownership', 'annual_inc', 'verification_status',
       'issue_d', 'loan_status', 'purpose', 'title', 'addr_state', 'dti',
       'delinq_2yrs', 'earliest_cr_line', 'inq_last_6mths', 'pub_rec',
       'total_acc', 'application_type', 'acc_open_past_24mths', 'delinq_amnt',
       'mort_acc', 'mths_since_recent_inq', 'num_bc_tl', 'num_il_tl',
       'num_op_rev_tl', 'pub_rec_bankruptcies', 'tax_liens', 'total_bc_limit',
       'total_il_high_credit_limit', 'average_fico'],
      dtype='object')
```

In [22]:
```
objects = list(df.loc[:,df.dtypes == 'object'].columns)
objects
```
executed in 158ms, finished 17:43:23 2021-04-05

Out[22]:
```
['term',
 'sub_grade',
 'home_ownership',
 'verification_status',
 'issue_d',
 'loan_status',
 'purpose',
 'title',
 'addr_state',
 'earliest_cr_line',
 'application_type']
```

In [48]:
```
categorical =['term',
 'sub_grade','grade',
 'home_ownership',
 'verification_status',
 'purpose','application_type','region']

drop = ['issue_d','earliest_cr_line']
cat_drop = ['term',
 'sub_grade',,'grade'
 'home_ownership',
 'verification_status',
 'purpose',
 'region','application_type','issue_d','earliest_cr_line']
cat_drop1 = ['term',
 'sub_grade',
 'home_ownership',
 'verification_status',
 'purpose',
 'region','application_type','issue_d','earliest_cr_line','loan_status']
```
executed in 18ms, finished 17:56:36 2021-04-05

In [25]:
```
df['earliest_cr_line'] = pd.to_datetime(df['earliest_cr_line'])
```
executed in 252ms, finished 17:46:31 2021-04-05

In [29]:
```
onehot = pd.get_dummies(df[categorical],drop_first=True)
```
executed in 1.49s, finished 17:47:34 2021-04-05

In [59]:
```
cont = df.drop(columns=cat_drop1)
```
executed in 75ms, finished 18:06:19 2021-04-05

In [62]:
```
encoded = pd.concat([cont,onehot],axis=1)
```
executed in 130ms, finished 18:18:59 2021-04-05

In [60]:
```
cont.describe()
```
executed in 1.41s, finished 18:06:24 2021-04-05

Out[60]:

	loan_amnt	int_rate	installment	emp_length	annual_inc	dti	delinq
count	1.736937e+06	1.736937e+06	1.736937e+06	1.736937e+06	1.736937e+06	1.736937e+06	1.736
mean	1.471386e+04	1.316545e+01	4.440634e+02	6.043523e+00	7.794385e+04	1.872237e+01	3.168
std	8.986702e+03	4.840233e+00	2.674402e+02	3.578266e+00	1.207552e+05	1.317191e+01	8.813
min	1.000000e+03	5.310000e+00	4.930000e+00	1.000000e+00	0.000000e+00	-1.000000e+00	0.000
25%	8.000000e+03	9.490000e+00	2.502000e+02	2.000000e+00	4.687200e+04	1.196000e+01	0.000
50%	1.200000e+04	1.269000e+01	3.768000e+02	6.000000e+00	6.500000e+04	1.786000e+01	0.000
75%	2.000000e+04	1.599000e+01	5.900000e+02	1.000000e+01	9.300000e+04	2.448000e+01	0.000
max	4.000000e+04	3.099000e+01	1.720000e+03	1.000000e+01	1.100000e+08	9.990000e+02	3.900

8 rows × 22 columns

In [120]:
```
na_check(df)
```
executed in 16.0s, finished 02:21:46 2021-04-07

Out[120]:
```
average_fico                 0
total_il_high_credit_limit   0
term                         0
int_rate                     0
installment                  0
sub_grade                    0
emp_length                   0
home_ownership               0
annual_inc                   0
verification_status          0
issue_d                      0
loan_status                  0
purpose                      0
addr_state                   0
dti                          0
delinq_2yrs                  0
earliest_cr_line             0
inq_last_6mths               0
pub_rec                      0
total_acc                    0
application_type             0
acc_open_past_24mths         0
delinq_amnt                  0
mort_acc                     0
mths_since_recent_inq        0
num_bc_tl                    0
num_il_tl                    0
num_op_rev_tl                0
pub_rec_bankruptcies         0
tax_liens                    0
total_bc_limit               0
loan_amnt                    0
dtype: int64
```

In [ ]:
```
smote= SMOTE()
x_smote,y_smote = smote.fit_sample(x_train,y_train)
x_train_sig =
```

In [ ]:

In [ ]:

In [40]:
```
x = encoded.drop(columns='loan_status',axis=1)
y = df['loan_status'].map(lambda x: 1 if x =="Charged Off" else 0)
x_train, x_test, y_train, y_test = train_test_split(x,y,test_size=.25)
```
executed in 1.78s, finished 17:51:22 2021-04-05

3 Sigmoid scaled no smote

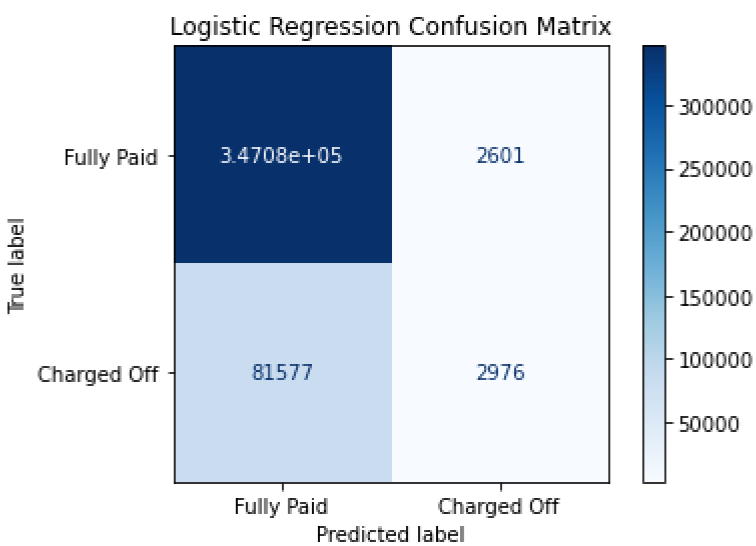
In [ ]:
```
x_train_sig = x_train
```

In [63]:
```
x_train, x_test, y_train, y_test = train_test_split(encoded,y,test_size=.25)
```
executed in 1.07s, finished 18:19:10 2021-04-05

In [98]:
```
x_train_sig = x_train.apply(sigmoid)
x_test_sig = x_test.apply(sigmoid)
```
executed in 5.81s, finished 02:52:54 2021-04-06

In [89]:
```
logreg = LogisticRegression(C=1,fit_intercept=True,solver='lbfgs')
log_model = logreg.fit(x_train_sig, y_train)
y_hat_log = logreg.predict(x_test_sig)
y_score_log = log_model.decision_function(x_test_sig)
fpr,tpr,thresholds = roc_curve(y_test,y_score_log)

print('AUC: {}'.format(auc(fpr, tpr)))
cf = confusion_matrix(y_test,y_hat_log)

plot_confusion_matrix(log_model,x_test_sig,y_test,cmap=plt.cm.Blues,
                     display_labels=["Fully Paid", "Charged Off"],
                      values_format=".5g")
plt.title("Logistic Regression Confusion Matrix")

print(confusion_matrix(y_test, y_hat_log))
print(classification_report(y_test, y_hat_log))
```
executed in 1m 51.2s, finished 02:39:09 2021-04-06

```
AUC: 0.7032559149668175
[[347081   2601]
 [ 81577   2976]]
              precision    recall  f1-score   support

           0       0.81      0.99      0.89    349682
           1       0.53      0.04      0.07     84553

    accuracy                           0.81    434235
   macro avg       0.67      0.51      0.48    434235
weighted avg       0.76      0.81      0.73    434235
```



In [42]:
```
from sklearn import datasets, linear_model
from sklearn.linear_model import LinearRegression
import statsmodels.api as sm
```
executed in 2.94s, finished 17:52:51 2021-04-05

In [97]:
```
def sigmoid(x):
    return 1/(1 + np.e**(-1*x))
```
executed in 7ms, finished 02:52:29 2021-04-06

In [ ]:

In [ ]:
```
result.
```

In [68]:
```python
logit_model=sm.Logit(y_train,x_train_sig)
result=logit_model.fit(method='bfgs',maxiter=100)
print(result.summary())
```

executed in 58.0s, finished 18:44:17 2021-04-05

```
Warning: Maximum number of iterations has been exceeded.
         Current function value: 0.453415
         Iterations: 100
         Function evaluations: 103
         Gradient evaluations: 103

C:\Users\sergi\anaconda3\envs\learn-env\lib\site-packages\statsmodels\base\mode
l.py:566: ConvergenceWarning: Maximum Likelihood optimization failed to converg
e. Check mle_retvals
  warnings.warn("Maximum Likelihood optimization failed to "

                           Logit Regression Results
==============================================================================
Dep. Variable:            loan_status   No. Observations:              1302702
Model:                          Logit   Df Residuals:                  1302574
Method:                           MLE   Df Model:                          127
Date:                Mon, 05 Apr 2021   Pseudo R-squ.:                 0.07800
Time:                        18:44:17   Log-Likelihood:            -5.9066e+05
converged:                      False   LL-Null:                   -6.4063e+05
Covariance Type:            nonrobust   LLR p-value:                     0.000
===============================================================================
========================
                                          coef    std err          z      P>|z|
[0.025      0.975]
-------------------------------------------------------------------------------
------------------------
loan_amnt                              -0.3091   2.08e+04  -1.48e-05      1.000
-4.08e+04    4.08e+04
int_rate                               -0.1012      9.684     -0.010      0.992
-19.082      18.880
installment                            -0.3091    283.583     -0.001      0.999
-556.122     555.504
emp_length                             -0.1750      0.024     -7.325      0.000
-0.222      -0.128
annual_inc                             -0.3067      2.914     -0.105      0.916
-6.017       5.404
dti                                     0.4046      0.088      4.578      0.000
0.231       0.578
delinq_2yrs                             0.4281      0.019     22.539      0.000
0.391       0.465
inq_last_6mths                          0.3838      0.017     22.545      0.000
0.350       0.417
pub_rec                                 0.3328      0.059      5.625      0.000
0.217       0.449
total_acc                              -0.3735      0.759     -0.492      0.623
-1.861       1.114
acc_open_past_24mths                    1.0875      0.024     45.478      0.000
1.041       1.134
delinq_amnt                            -0.0279      0.075     -0.370      0.711
-0.176       0.120
mort_acc                               -0.8179      0.015    -53.614      0.000
-0.848      -0.788
```

```
mths_since_recent_inq                    -0.2606      0.016    -15.986      0.000
-0.293      -0.229
num_bc_tl                                -0.5236      0.055     -9.466      0.000
-0.632      -0.415
num_il_tl                                -0.3169      0.028    -11.267      0.000
-0.372      -0.262
num_op_rev_tl                             0.3524      0.087      4.058      0.000
0.182       0.523
pub_rec_bankruptcies                      0.2409      0.063      3.833      0.000
0.118       0.364
tax_liens                                -0.1315      0.068     -1.932      0.053
-0.265       0.002
total_bc_limit                           -0.2414      0.305     -0.792      0.428
-0.839       0.356
total_il_high_credit_limit               -0.0104      0.019     -0.558      0.577
-0.047       0.026
average_fico                             -0.3091   2.08e+04  -1.48e-05      1.000
-4.08e+04    4.08e+04
term_ 60 months                           2.8556      0.024    118.702      0.000
2.808       2.903
sub_grade_A2                             -4.0739      0.152    -26.828      0.000
-4.372      -3.776
sub_grade_A3                             -3.8641      0.168    -22.948      0.000
-4.194      -3.534
sub_grade_A4                             -3.9743      0.173    -23.012      0.000
-4.313      -3.636
sub_grade_A5                             -3.5984      0.178    -20.201      0.000
-3.948      -3.249
sub_grade_B1                             -2.9126      0.180    -16.181      0.000
-3.265      -2.560
sub_grade_B2                             -2.4938      0.183    -13.624      0.000
-2.853      -2.135
sub_grade_B3                             -1.9453      0.183    -10.627      0.000
-2.304      -1.587
sub_grade_B4                             -1.3821      0.183     -7.567      0.000
-1.740      -1.024
sub_grade_B5                             -0.8946      0.182     -4.903      0.000
-1.252      -0.537
sub_grade_C1                             -0.3316      0.182     -1.821      0.069
-0.688       0.025
sub_grade_C2                              0.0618      0.182      0.339      0.734
-0.295       0.419
sub_grade_C3                              0.2973      0.182      1.631      0.103
-0.060       0.655
sub_grade_C4                              0.7029      0.182      3.858      0.000
0.346       1.060
sub_grade_C5                              0.9631      0.182      5.279      0.000
0.605       1.321
sub_grade_D1                              1.3163      0.184      7.161      0.000
0.956       1.677
sub_grade_D2                              1.8309      0.184      9.947      0.000
1.470       2.192
sub_grade_D3                              1.9976      0.185     10.797      0.000
1.635       2.360
sub_grade_D4                              2.1132      0.186     11.375      0.000
1.749       2.477
sub_grade_D5                              2.1999      0.187     11.760      0.000
```

```
1.833       2.567
sub_grade_E1                              2.0787      0.190     10.926      0.000
1.706       2.452
sub_grade_E2                              2.1129      0.192     11.031      0.000
1.738       2.488
sub_grade_E3                              2.2198      0.193     11.487      0.000
1.841       2.599
sub_grade_E4                              2.0288      0.196     10.366      0.000
1.645       2.412
sub_grade_E5                              2.1883      0.196     11.156      0.000
1.804       2.573
sub_grade_F1                              1.2979      0.207      6.268      0.000
0.892       1.704
sub_grade_F2                              1.2100      0.218      5.552      0.000
0.783       1.637
sub_grade_F3                              0.9438      0.225      4.197      0.000
0.503       1.385
sub_grade_F4                              0.9387      0.235      4.000      0.000
0.479       1.399
sub_grade_F5                              0.7934      0.246      3.231      0.001
0.312       1.275
sub_grade_G1                              0.5137      0.262      1.963      0.050
0.001       1.027
sub_grade_G2                              0.3421      0.294      1.162      0.245
-0.235       0.919
sub_grade_G3                              0.2576      0.320      0.804      0.421
-0.370       0.886
sub_grade_G4                              0.2077      0.348      0.598      0.550
-0.473       0.889
sub_grade_G5                              0.2258      0.358      0.630      0.529
-0.477       0.928
home_ownership_MORTGAGE                  -0.6925      0.392     -1.766      0.077
-1.461       0.076
home_ownership_NONE                      -0.1599      2.073     -0.077      0.939
-4.224       3.904
home_ownership_OTHER                     -0.1583      2.163     -0.073      0.942
-4.399       4.082
home_ownership_OWN                       -0.1394      0.393     -0.355      0.723
-0.909       0.631
home_ownership_RENT                       0.2874      0.392      0.733      0.464
-0.481       1.056
verification_status_Source Verified       0.3174      0.025     12.484      0.000
0.268       0.367
verification_status_Verified              0.7856      0.027     28.993      0.000
0.733       0.839
purpose_credit_card                      -0.0295      0.106     -0.279      0.781
-0.237       0.178
purpose_debt_consolidation               -0.0303      0.104     -0.291      0.771
-0.235       0.174
purpose_educational                      -0.1554      7.240     -0.021      0.983
-14.346      14.036
purpose_home_improvement                 -0.3414      0.112     -3.057      0.002
-0.560      -0.122
purpose_house                            -0.1588      0.163     -0.976      0.329
-0.477       0.160
purpose_major_purchase                   -0.3302      0.125     -2.650      0.008
-0.574      -0.086
```

purpose_medical	0.1288	0.137	0.940	0.347	-0.140	0.397
purpose_moving	-0.1298	0.154	-0.845	0.398	-0.431	0.171
purpose_other	-0.4769	0.111	-4.286	0.000	-0.695	-0.259
purpose_renewable_energy	-0.1608	0.397	-0.405	0.686	-0.939	0.618
purpose_small_business	1.1275	0.136	8.263	0.000	0.860	1.395
purpose_vacation	-0.2752	0.159	-1.731	0.084	-0.587	0.036
purpose_wedding	-0.2516	0.430	-0.585	0.558	-1.094	0.591
addr_state_AL	0.7089	0.225	3.152	0.002	0.268	1.150
addr_state_AR	0.6933	0.235	2.953	0.003	0.233	1.154
addr_state_AZ	-0.0369	0.217	-0.170	0.865	-0.463	0.389
addr_state_CA	-0.2138	0.209	-1.021	0.307	-0.624	0.196
addr_state_CO	-1.4164	0.221	-6.423	0.000	-1.849	-0.984
addr_state_CT	-0.5021	0.224	-2.241	0.025	-0.941	-0.063
addr_state_DC	-0.4708	0.296	-1.588	0.112	-1.052	0.110
addr_state_DE	-0.1803	0.280	-0.643	0.520	-0.730	0.369
addr_state_FL	0.4171	0.211	1.979	0.048	0.004	0.830
addr_state_GA	-0.4125	0.215	-1.916	0.055	-0.835	0.010
addr_state_HI	-0.2447	0.251	-0.977	0.329	-0.736	0.246
addr_state_IA	-0.1551	9.825	-0.016	0.987	-19.413	19.102
addr_state_ID	-0.3002	0.316	-0.950	0.342	-0.919	0.319
addr_state_IL	-0.6288	0.214	-2.936	0.003	-1.049	-0.209
addr_state_IN	0.3039	0.221	1.376	0.169	-0.129	0.737
addr_state_KS	-0.5232	0.237	-2.206	0.027	-0.988	-0.058
addr_state_KY	0.0637	0.231	0.276	0.783	-0.389	0.517
addr_state_LA	0.5721	0.227	2.523	0.012	0.128	1.017
addr_state_MA	-0.3198	0.218	-1.465	0.143	-0.748	0.108
addr_state_MD	0.3229	0.217	1.486	0.137	-0.103	0.749
addr_state_ME	-0.4258	0.311	-1.368	0.171	-1.036	0.184
addr_state_MI	-0.0108	0.217	-0.050	0.960		

					-0.435	0.414
addr_state_MN	-0.1385	0.221	-0.627	0.531	-0.572	0.295
addr_state_MO	0.2634	0.222	1.187	0.235	-0.172	0.699
addr_state_MS	0.5281	0.245	2.155	0.031	0.048	1.008
addr_state_MT	-0.3370	0.284	-1.186	0.236	-0.894	0.220
addr_state_NC	0.2636	0.216	1.222	0.222	-0.159	0.687
addr_state_ND	-0.2166	0.331	-0.655	0.513	-0.865	0.432
addr_state_NE	0.0827	0.270	0.307	0.759	-0.446	0.611
addr_state_NH	-0.5832	0.256	-2.280	0.023	-1.084	-0.082
addr_state_NJ	-0.0387	0.214	-0.181	0.856	-0.458	0.381
addr_state_NM	0.0311	0.249	0.125	0.900	-0.456	0.518
addr_state_NV	0.3586	0.222	1.613	0.107	-0.077	0.794
addr_state_NY	0.1889	0.210	0.898	0.369	-0.223	0.601
addr_state_OH	-0.1018	0.215	-0.474	0.636	-0.523	0.319
addr_state_OK	0.5214	0.231	2.257	0.024	0.069	0.974
addr_state_OR	-1.3098	0.230	-5.694	0.000	-1.761	-0.859
addr_state_PA	0.0709	0.214	0.330	0.741	-0.349	0.491
addr_state_RI	-0.3338	0.258	-1.295	0.195	-0.839	0.172
addr_state_SC	-0.9215	0.229	-4.020	0.000	-1.371	-0.472
addr_state_SD	-0.1012	0.306	-0.331	0.741	-0.701	0.499
addr_state_TN	0.0558	0.222	0.252	0.801	-0.379	0.490
addr_state_TX	0.0971	0.211	0.461	0.644	-0.315	0.510
addr_state_UT	-0.4824	0.241	-2.002	0.045	-0.955	-0.010
addr_state_VA	0.0688	0.216	0.318	0.750	-0.355	0.492
addr_state_VT	-0.4393	0.309	-1.424	0.155	-1.044	0.165
addr_state_WA	-1.5452	0.221	-6.995	0.000	-1.978	-1.112
addr_state_WI	-0.2749	0.226	-1.218	0.223	-0.717	0.167
addr_state_WV	-0.4619	0.270	-1.710	0.087	-0.991	0.067
addr_state_WY	-0.2532	0.304	-0.832	0.406	-0.850	0.343

```
application_type_Joint App              1.7688      0.049     36.183      0.000
1.673       1.865
================================================================================
========================
```

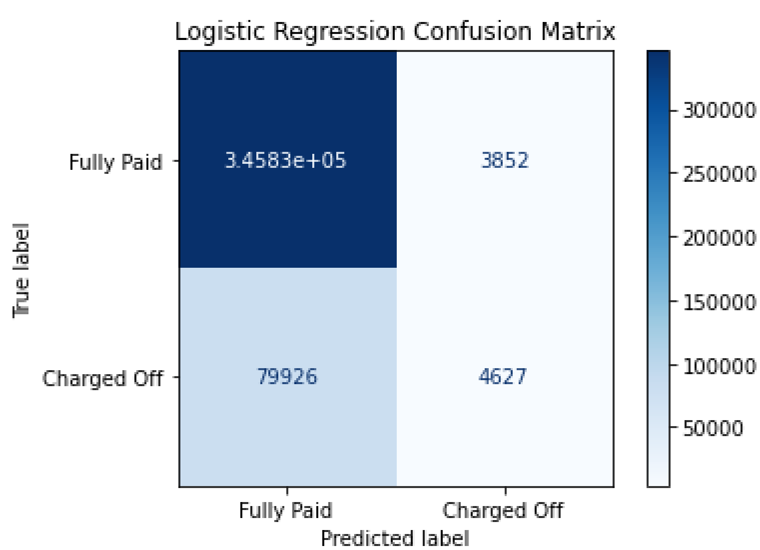
4 Min Max scaled no smote

In [69]:
```
scaler = StandardScaler()
x_train_scaled = scaler.fit_transform(x_train)
x_test_scaled = scaler.transform(x_test)
```
executed in 5.25s, finished 02:19:09 2021-04-06

In [90]:
```python
logreg = LogisticRegression(C=1,fit_intercept=True,solver='lbfgs')
log_model = logreg.fit(x_train_scaled, y_train)
y_hat_log = logreg.predict(x_test_scaled)
y_score_log = log_model.decision_function(x_test_scaled)
fpr,tpr,thresholds = roc_curve(y_test,y_score_log)

print('AUC: {}'.format(auc(fpr, tpr)))
cf = confusion_matrix(y_test,y_hat_log)

plot_confusion_matrix(log_model,x_test_scaled,y_test,cmap=plt.cm.Blues,
                      display_labels=["Fully Paid", "Charged Off"],
                       values_format=".5g")
plt.title("Logistic Regression Confusion Matrix")

print(confusion_matrix(y_test, y_hat_log))
print(classification_report(y_test, y_hat_log))
```

executed in 41.9s, finished 02:40:25 2021-04-06

```
AUC: 0.711299165599099
[[345830   3852]
 [ 79926   4627]]
              precision    recall  f1-score   support

           0       0.81      0.99      0.89    349682
           1       0.55      0.05      0.10     84553

    accuracy                           0.81    434235
   macro avg       0.68      0.52      0.50    434235
weighted avg       0.76      0.81      0.74    434235
```



In [ ]:

In [ ]:

```
logreg = LogisticRegression(C=1e12,fit_intercept=True,solver='lbfgs')
log_model = logreg.fit(x_train_scaled, y_train)
y_hat_log = logreg.predict(x_test_scaled)
y_score_log = log_model.decision_function(x_test_scaled)
fpr,tpr,thresholds = roc_curve(y_test,y_score_log)

print('AUC: {}'.format(auc(fpr, tpr)))
cf = confusion_matrix(y_test,y_hat_log)

plot_confusion_matrix(log_model,x_test_scaled,y_test,cmap=plt.cm.Blues,
                     display_labels=["Fully Paid", "Charged Off"],
                      values_format=".5g")
plt.title("Logistic Regression Confusion Matrix")

print(confusion_matrix(y_test, y_hat_log))
print(classification_report(y_test, y_hat_log))
```

In [78]:

```
pd.options.display.max_rows = 128
```

executed in 12ms, finished 02:30:09 2021-04-06

In [87]:
```
scaled_results = pd.DataFrame(zip(x_train.columns,result.params.values,np.round(r
scaled_results
```
executed in 31ms, finished 02:32:16 2021-04-06

Out[87]:

	features	coef	pvalues
0	loan_amnt	-0.115929	0.0000
1	int_rate	0.059899	0.0000
2	installment	0.198262	0.0000
3	emp_length	-0.015526	0.0000
4	annual_inc	-0.024591	0.0000
5	dti	0.084440	0.0000
6	delinq_2yrs	0.023995	0.0000
7	inq_last_6mths	0.016792	0.0000
8	pub_rec	-0.003348	0.5359
9	total_acc	-0.137110	0.0000
10	acc_open_past_24mths	0.110534	0.0000
11	delinq_amnt	0.002220	0.2180
12	mort_acc	-0.043998	0.0000
13	mths_since_recent_inq	-0.022521	0.0000
14	num_bc_tl	0.027251	0.0000
15	num_il_tl	0.049885	0.0000
16	num_op_rev_tl	0.066243	0.0000
17	pub_rec_bankruptcies	0.013387	0.0005
18	tax_liens	0.004854	0.2357
19	total_bc_limit	-0.060604	0.0000
20	total_il_high_credit_limit	-0.035257	0.0000
21	average_fico	-0.050546	0.0000
22	term_ 60 months	0.221475	0.0000
23	sub_grade_A2	-0.004250	0.0862
24	sub_grade_A3	-0.006989	0.0057
25	sub_grade_A4	-0.007670	0.0053
26	sub_grade_A5	-0.006444	0.0305
27	sub_grade_B1	-0.002453	0.4543
28	sub_grade_B2	-0.001759	0.6143
29	sub_grade_B3	0.002530	0.5015

	features	coef	pvalues
30	sub_grade_B4	0.009716	0.0168
31	sub_grade_B5	0.017678	0.0000
32	sub_grade_C1	0.027996	0.0000
33	sub_grade_C2	0.035800	0.0000
34	sub_grade_C3	0.041162	0.0000
35	sub_grade_C4	0.052709	0.0000
36	sub_grade_C5	0.055641	0.0000
37	sub_grade_D1	0.051548	0.0000
38	sub_grade_D2	0.058474	0.0000
39	sub_grade_D3	0.057924	0.0000
40	sub_grade_D4	0.058677	0.0000
41	sub_grade_D5	0.056850	0.0000
42	sub_grade_E1	0.052492	0.0000
43	sub_grade_E2	0.053532	0.0000
44	sub_grade_E3	0.057384	0.0000
45	sub_grade_E4	0.054151	0.0000
46	sub_grade_E5	0.057424	0.0000
47	sub_grade_F1	0.039981	0.0000
48	sub_grade_F2	0.043683	0.0000
49	sub_grade_F3	0.035917	0.0000
50	sub_grade_F4	0.041178	0.0000
51	sub_grade_F5	0.037577	0.0000
52	sub_grade_G1	0.028133	0.0000
53	sub_grade_G2	0.025986	0.0000
54	sub_grade_G3	0.025395	0.0000
55	sub_grade_G4	0.024913	0.0000
56	sub_grade_G5	0.028174	0.0000
57	home_ownership_MORTGAGE	-0.052616	0.1358
58	home_ownership_NONE	-0.002729	0.1482
59	home_ownership_OTHER	-0.001803	0.3281
60	home_ownership_OWN	-0.002812	0.8994
61	home_ownership_RENT	0.037416	0.2774
62	verification_status_Source Verified	0.008793	0.0001

	features	coef	pvalues
63	verification_status_Verified	0.014901	0.0000
64	purpose_credit_card	0.020881	0.0071
65	purpose_debt_consolidation	0.028570	0.0016
66	purpose_educational	-0.000935	0.6310
67	purpose_home_improvement	0.021880	0.0000
68	purpose_house	0.004320	0.0581
69	purpose_major_purchase	0.015973	0.0000
70	purpose_medical	0.016782	0.0000
71	purpose_moving	0.009314	0.0001
72	purpose_other	0.017704	0.0001
73	purpose_renewable_energy	0.000981	0.5975
74	purpose_small_business	0.030309	0.0000
75	purpose_vacation	0.006642	0.0045
76	purpose_wedding	-0.007781	0.0000
77	addr_state_AL	0.016656	0.0002
78	addr_state_AR	0.019382	0.0000
79	addr_state_AZ	0.003029	0.6176
80	addr_state_CA	-0.002455	0.8521
81	addr_state_CO	-0.020732	0.0003
82	addr_state_CT	-0.003148	0.5204
83	addr_state_DC	-0.007725	0.0026
84	addr_state_DE	0.000102	0.9697
85	addr_state_FL	0.019695	0.0450
86	addr_state_GA	-0.004596	0.5040
87	addr_state_HI	-0.002686	0.3952
88	addr_state_IA	-0.000888	0.6410
89	addr_state_ID	-0.006495	0.0072
90	addr_state_IL	-0.004999	0.5037
91	addr_state_IN	0.009614	0.0637
92	addr_state_KS	-0.006174	0.1082
93	addr_state_KY	0.005861	0.1509
94	addr_state_LA	0.015380	0.0004
95	addr_state_MA	0.001757	0.7652

	features	coef	pvalues
96	addr_state_MD	0.011810	0.0460
97	addr_state_ME	-0.009065	0.0003
98	addr_state_MI	0.005279	0.3993
99	addr_state_MN	0.002366	0.6543
100	addr_state_MO	0.009487	0.0590
101	addr_state_MS	0.016775	0.0000
102	addr_state_MT	-0.005130	0.0566
103	addr_state_NC	0.008163	0.2049
104	addr_state_ND	-0.003591	0.1212
105	addr_state_NE	0.007264	0.0093
106	addr_state_NH	-0.009125	0.0043
107	addr_state_NJ	0.007844	0.2754
108	addr_state_NM	0.005031	0.1249
109	addr_state_NV	0.008028	0.1008
110	addr_state_NY	0.019981	0.0535
111	addr_state_OH	0.002279	0.7419
112	addr_state_OK	0.014867	0.0002
113	addr_state_OR	-0.020526	0.0000
114	addr_state_PA	0.007453	0.2856
115	addr_state_RI	-0.003689	0.2288
116	addr_state_SC	-0.015482	0.0005
117	addr_state_SD	0.002160	0.3803
118	addr_state_TN	0.003329	0.5110
119	addr_state_TX	0.007620	0.4665
120	addr_state_UT	-0.007321	0.0446
121	addr_state_VA	0.006026	0.3461
122	addr_state_VT	-0.009894	0.0001
123	addr_state_WA	-0.023893	0.0000
124	addr_state_WI	0.001874	0.6880
125	addr_state_WV	-0.008283	0.0038
126	addr_state_WY	-0.003760	0.1327
127	application_type_Joint App	0.019906	0.0000

In [82]:
```
result.pvalues.values
```
executed in 13ms, finished 02:30:43 2021-04-06

Out[82]:
```
array([2.53280741e-016, 1.60032929e-011, 3.95619674e-051, 1.23391324e-016,
       1.80514515e-008, 2.43561337e-183, 7.58337735e-037, 4.33796465e-014,
       5.35947843e-001, 5.37520012e-095, 0.00000000e+000, 2.18035606e-001,
       2.36459041e-062, 3.74896536e-026, 2.21161158e-013, 3.74044655e-025,
       2.62874113e-088, 5.05450747e-004, 2.35708951e-001, 4.13959711e-112,
       8.17947427e-045, 1.60497734e-099, 0.00000000e+000, 8.62288758e-002,
       5.70968704e-003, 5.31558492e-003, 3.05388587e-002, 4.54290884e-001,
       6.14325784e-001, 5.01531870e-001, 1.68441450e-002, 3.15502886e-005,
       8.16976959e-010, 1.07747431e-014, 6.85297202e-018, 9.44487251e-026,
       3.36018693e-027, 2.54882038e-027, 2.78854095e-033, 2.67543432e-034,
       4.78016900e-036, 7.11246811e-035, 8.14104180e-038, 5.01254886e-040,
       5.01457604e-047, 7.84764003e-044, 1.09934279e-045, 1.12663332e-030,
       2.16122322e-042, 4.52326874e-031, 3.65458096e-042, 3.64787423e-037,
       7.25253208e-024, 5.57797630e-023, 2.12562383e-022, 1.74319632e-021,
       3.50543225e-025, 1.35784225e-001, 1.48155505e-001, 3.28060796e-001,
       8.99408237e-001, 2.77353475e-001, 5.34687880e-005, 3.26222317e-011,
       7.08110590e-003, 1.58257963e-003, 6.31044617e-001, 5.95637379e-006,
       5.81178860e-002, 4.87430394e-007, 2.54682533e-010, 6.52321623e-005,
       1.26951828e-004, 5.97489514e-001, 2.98291824e-032, 4.51523968e-003,
       2.81260687e-005, 1.99516773e-004, 1.98225501e-007, 6.17568883e-001,
       8.52097906e-001, 3.34774187e-004, 5.20393150e-001, 2.64802167e-003,
       9.69672199e-001, 4.50272520e-002, 5.04035186e-001, 3.95175130e-001,
       6.40953869e-001, 7.15122263e-003, 5.03696251e-001, 6.37098786e-002,
       1.08194070e-001, 1.50939453e-001, 3.97474437e-004, 7.65227239e-001,
       4.59923279e-002, 2.50167329e-004, 3.99295172e-001, 6.54321897e-001,
       5.89919506e-002, 3.59264777e-007, 5.66212880e-002, 2.04921355e-001,
       1.21246472e-001, 9.32858428e-003, 4.26927676e-003, 2.75378165e-001,
       1.24900734e-001, 1.00840193e-001, 5.35429261e-002, 7.41873212e-001,
       1.98159966e-004, 4.47999246e-006, 2.85633303e-001, 2.28810608e-001,
       5.47095259e-004, 3.80326691e-001, 5.11017120e-001, 4.66485064e-001,
       4.45505044e-002, 3.46133072e-001, 7.48023661e-005, 2.78777166e-005,
       6.88032666e-001, 3.75999894e-003, 1.32726911e-001, 3.07801953e-025])
```

In [73]:
```
logit_model=sm.Logit(y_train,x_train_scaled)
result=logit_model.fit(method='bfgs',maxiter=100)
print(result.summary())
```
executed in 1m 24.0s, finished 02:25:44 2021-04-06

```
Optimization terminated successfully.
         Current function value: 0.664410
         Iterations: 82
         Function evaluations: 83
         Gradient evaluations: 83
                           Logit Regression Results
==============================================================================
Dep. Variable:            loan_status   No. Observations:              1302702
Model:                          Logit   Df Residuals:                  1302574
Method:                           MLE   Df Model:                          127
Date:                Tue, 06 Apr 2021   Pseudo R-squ.:                 -0.3511
Time:                        02:25:43   Log-Likelihood:            -8.6553e+05
converged:                       True   LL-Null:                   -6.4063e+05
Covariance Type:            nonrobust   LLR p-value:                     1.000
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
x1            -0.1159      0.014     -8.194      0.000      -0.144      -0.088
x2             0.0599      0.009      6.738      0.000       0.042       0.077
x3             0.1983      0.013     15.041      0.000       0.172       0.224
x4            -0.0155      0.002     -8.280      0.000      -0.019      -0.012
x5            -0.0246      0.004     -5.630      0.000      -0.033      -0.016
x6             0.0844      0.003     28.875      0.000       0.079       0.090
x7             0.0240      0.002     12.681      0.000       0.020       0.028
x8             0.0168      0.002      7.550      0.000       0.012       0.021
x9            -0.0033      0.005     -0.619      0.536      -0.014       0.007
x10           -0.1371      0.007    -20.679      0.000      -0.150      -0.124
x11            0.1105      0.002     47.681      0.000       0.106       0.115
x12            0.0022      0.002      1.232      0.218      -0.001       0.006
x13           -0.0440      0.003    -16.665      0.000      -0.049      -0.039
x14           -0.0225      0.002    -10.579      0.000      -0.027      -0.018
x15            0.0273      0.004      7.335      0.000       0.020       0.035
x16            0.0499      0.005     10.361      0.000       0.040       0.059
x17            0.0662      0.003     19.922      0.000       0.060       0.073
x18            0.0134      0.004      3.478      0.001       0.006       0.021
x19            0.0049      0.004      1.186      0.236      -0.003       0.013
x20           -0.0606      0.003    -22.500      0.000      -0.066      -0.055
x21           -0.0353      0.003    -14.046      0.000      -0.040      -0.030
x22           -0.0505      0.002    -21.176      0.000      -0.055      -0.046
x23            0.2215      0.005     46.482      0.000       0.212       0.231
x24           -0.0043      0.002     -1.716      0.086      -0.009       0.001
x25           -0.0070      0.003     -2.764      0.006      -0.012      -0.002
x26           -0.0077      0.003     -2.787      0.005      -0.013      -0.002
x27           -0.0064      0.003     -2.163      0.031      -0.012      -0.001
x28           -0.0025      0.003     -0.748      0.454      -0.009       0.004
x29           -0.0018      0.003     -0.504      0.614      -0.009       0.005
x30            0.0025      0.004      0.672      0.502      -0.005       0.010
x31            0.0097      0.004      2.390      0.017       0.002       0.018
x32            0.0177      0.004      4.162      0.000       0.009       0.026
x33            0.0280      0.005      6.142      0.000       0.019       0.037
x34            0.0358      0.005      7.730      0.000       0.027       0.045
x35            0.0412      0.005      8.617      0.000       0.032       0.051
```

x36	0.0527	0.005	10.492	0.000	0.043	0.063
x37	0.0556	0.005	10.802	0.000	0.046	0.066
x38	0.0515	0.005	10.828	0.000	0.042	0.061
x39	0.0585	0.005	12.020	0.000	0.049	0.068
x40	0.0579	0.005	12.212	0.000	0.049	0.067
x41	0.0587	0.005	12.535	0.000	0.050	0.068
x42	0.0568	0.005	12.320	0.000	0.048	0.066
x43	0.0525	0.004	12.854	0.000	0.044	0.060
x44	0.0535	0.004	13.242	0.000	0.046	0.061
x45	0.0574	0.004	14.402	0.000	0.050	0.065
x46	0.0542	0.004	13.885	0.000	0.047	0.062
x47	0.0574	0.004	14.187	0.000	0.049	0.065
x48	0.0400	0.003	11.514	0.000	0.033	0.047
x49	0.0437	0.003	13.645	0.000	0.037	0.050
x50	0.0359	0.003	11.592	0.000	0.030	0.042
x51	0.0412	0.003	13.607	0.000	0.035	0.047
x52	0.0376	0.003	12.738	0.000	0.032	0.043
x53	0.0281	0.003	10.073	0.000	0.023	0.034
x54	0.0260	0.003	9.871	0.000	0.021	0.031
x55	0.0254	0.003	9.736	0.000	0.020	0.031
x56	0.0249	0.003	9.519	0.000	0.020	0.030
x57	0.0282	0.003	10.367	0.000	0.023	0.034
x58	-0.0526	0.035	-1.492	0.136	-0.122	0.017
x59	-0.0027	0.002	-1.446	0.148	-0.006	0.001
x60	-0.0018	0.002	-0.978	0.328	-0.005	0.002
x61	-0.0028	0.022	-0.126	0.899	-0.046	0.041
x62	0.0374	0.034	1.086	0.277	-0.030	0.105
x63	0.0088	0.002	4.040	0.000	0.005	0.013
x64	0.0149	0.002	6.634	0.000	0.010	0.019
x65	0.0209	0.008	2.693	0.007	0.006	0.036
x66	0.0286	0.009	3.159	0.002	0.011	0.046
x67	-0.0009	0.002	-0.480	0.631	-0.005	0.003
x68	0.0219	0.005	4.528	0.000	0.012	0.031
x69	0.0043	0.002	1.895	0.058	-0.000	0.009
x70	0.0160	0.003	5.031	0.000	0.010	0.022
x71	0.0168	0.003	6.324	0.000	0.012	0.022
x72	0.0093	0.002	3.993	0.000	0.005	0.014
x73	0.0177	0.005	3.832	0.000	0.009	0.027
x74	0.0010	0.002	0.528	0.597	-0.003	0.005
x75	0.0303	0.003	11.823	0.000	0.025	0.035
x76	0.0066	0.002	2.840	0.005	0.002	0.011
x77	-0.0078	0.002	-4.188	0.000	-0.011	-0.004
x78	0.0167	0.004	3.720	0.000	0.008	0.025
x79	0.0194	0.004	5.201	0.000	0.012	0.027
x80	0.0030	0.006	0.499	0.618	-0.009	0.015
x81	-0.0025	0.013	-0.186	0.852	-0.028	0.023
x82	-0.0207	0.006	-3.587	0.000	-0.032	-0.009
x83	-0.0031	0.005	-0.643	0.520	-0.013	0.006
x84	-0.0077	0.003	-3.006	0.003	-0.013	-0.003
x85	0.0001	0.003	0.038	0.970	-0.005	0.005
x86	0.0197	0.010	2.004	0.045	0.000	0.039
x87	-0.0046	0.007	-0.668	0.504	-0.018	0.009
x88	-0.0027	0.003	-0.850	0.395	-0.009	0.004
x89	-0.0009	0.002	-0.466	0.641	-0.005	0.003
x90	-0.0065	0.002	-2.690	0.007	-0.011	-0.002
x91	-0.0050	0.007	-0.669	0.504	-0.020	0.010
x92	0.0096	0.005	1.854	0.064	-0.001	0.020

```
x93           -0.0062      0.004     -1.606      0.108      -0.014       0.001
x94            0.0059      0.004      1.436      0.151      -0.002       0.014
x95            0.0154      0.004      3.542      0.000       0.007       0.024
x96            0.0018      0.006      0.299      0.765      -0.010       0.013
x97            0.0118      0.006      1.995      0.046       0.000       0.023
x98           -0.0091      0.002     -3.662      0.000      -0.014      -0.004
x99            0.0053      0.006      0.843      0.399      -0.007       0.018
x100           0.0024      0.005      0.448      0.654      -0.008       0.013
x101           0.0095      0.005      1.888      0.059      -0.000       0.019
x102           0.0168      0.003      5.089      0.000       0.010       0.023
x103          -0.0051      0.003     -1.906      0.057      -0.010       0.000
x104           0.0082      0.006      1.268      0.205      -0.004       0.021
x105          -0.0036      0.002     -1.550      0.121      -0.008       0.001
x106           0.0073      0.003      2.600      0.009       0.002       0.013
x107          -0.0091      0.003     -2.858      0.004      -0.015      -0.003
x108           0.0078      0.007      1.091      0.275      -0.006       0.022
x109           0.0050      0.003      1.535      0.125      -0.001       0.011
x110           0.0080      0.005      1.641      0.101      -0.002       0.018
x111           0.0200      0.010      1.931      0.054      -0.000       0.040
x112           0.0023      0.007      0.329      0.742      -0.011       0.016
x113           0.0149      0.004      3.721      0.000       0.007       0.023
x114          -0.0205      0.004     -4.588      0.000      -0.029      -0.012
x115           0.0075      0.007      1.068      0.286      -0.006       0.021
x116          -0.0037      0.003     -1.203      0.229      -0.010       0.002
x117          -0.0155      0.004     -3.457      0.001      -0.024      -0.007
x118           0.0022      0.002      0.877      0.380      -0.003       0.007
x119           0.0033      0.005      0.657      0.511      -0.007       0.013
x120           0.0076      0.010      0.728      0.466      -0.013       0.028
x121          -0.0073      0.004     -2.009      0.045      -0.014      -0.000
x122           0.0060      0.006      0.942      0.346      -0.007       0.019
x123          -0.0099      0.002     -3.960      0.000      -0.015      -0.005
x124          -0.0239      0.006     -4.190      0.000      -0.035      -0.013
x125           0.0019      0.005      0.402      0.688      -0.007       0.011
x126          -0.0083      0.003     -2.898      0.004      -0.014      -0.003
x127          -0.0038      0.003     -1.503      0.133      -0.009       0.001
x128           0.0199      0.002     10.379      0.000       0.016       0.024
================================================================================
```

In [ ]:
```
smote= SMOTE()
x_smote,y_smote = smote.fit_sample(x_train,y_train)
```

In [99]:
```


x_train_smsig = x_smote.apply(sigmoid)

x_train_smscaled = scaler.fit_transform(x_smote)
```

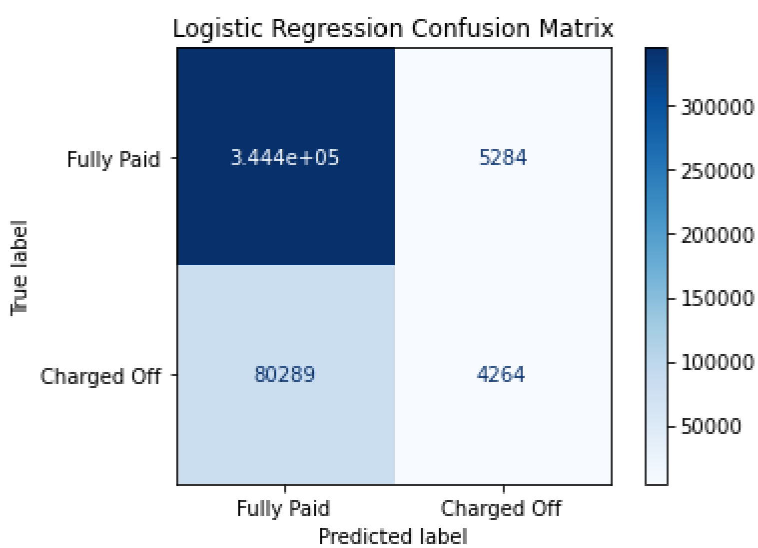
executed in 22.7s, finished 02:53:51 2021-04-06

5 Log Smote

In [100]:
```python
logreg = LogisticRegression(C=1e7,fit_intercept=True,solver='sag')
log_model = logreg.fit(x_train_smsig, y_smote)
y_hat_log = logreg.predict(x_test_sig)
y_score_log = log_model.decision_function(x_test_sig)
fpr,tpr,thresholds = roc_curve(y_test,y_score_log)

print('AUC: {}'.format(auc(fpr, tpr)))
cf = confusion_matrix(y_test,y_hat_log)

plot_confusion_matrix(log_model,x_test_sig,y_test,cmap=plt.cm.Blues,
                     display_labels=["Fully Paid", "Charged Off"],
                      values_format=".5g")
plt.title("Logistic Regression Confusion Matrix")

print(confusion_matrix(y_test, y_hat_log))
print(classification_report(y_test, y_hat_log))
```

executed in 4m 46s, finished 03:00:10 2021-04-06

```
C:\Users\sergi\anaconda3\envs\learn-env\lib\site-packages\sklearn\linear_model
\_sag.py:329: ConvergenceWarning: The max_iter was reached which means the coef
_ did not converge
  warnings.warn("The max_iter was reached which means "

AUC: 0.6888340766512712
[[344398   5284]
 [ 80289   4264]]
              precision    recall  f1-score   support

           0       0.81      0.98      0.89    349682
           1       0.45      0.05      0.09     84553

    accuracy                           0.80    434235
   macro avg       0.63      0.52      0.49    434235
weighted avg       0.74      0.80      0.73    434235
```



In [105]:
```
logit_model=sm.Logit(y_smote,x_train_smsig)
result=logit_model.fit(method='bfgs',maxiter=100)
print(result.summary())
```
executed in 1m 37.9s, finished 03:31:24 2021-04-06

```
Warning: Maximum number of iterations has been exceeded.
         Current function value: 0.365074
         Iterations: 100
         Function evaluations: 103
         Gradient evaluations: 103

C:\Users\sergi\anaconda3\envs\learn-env\lib\site-packages\statsmodels\base\mode
l.py:547: HessianInversionWarning: Inverting hessian failed, no bse or cov_para
ms available
  warnings.warn('Inverting hessian failed, no bse or cov_params '
C:\Users\sergi\anaconda3\envs\learn-env\lib\site-packages\statsmodels\base\mode
l.py:566: ConvergenceWarning: Maximum Likelihood optimization failed to converg
e. Check mle_retvals
  warnings.warn("Maximum Likelihood optimization failed to "

                           Logit Regression Results
==============================================================================
Dep. Variable:            loan_status   No. Observations:              2100320
Model:                          Logit   Df Residuals:                  2100192
Method:                           MLE   Df Model:                          127
Date:                Tue, 06 Apr 2021   Pseudo R-squ.:                  0.4733
Time:                        03:31:23   Log-Likelihood:            -7.6677e+05
converged:                      False   LL-Null:                   -1.4558e+06
Covariance Type:            nonrobust   LLR p-value:                     0.000
================================================================================
========================
                                          coef    std err          z      P>|z|
[0.025      0.975]
--------------------------------------------------------------------------------
------------------------
loan_amnt                               8.6440        nan        nan        nan
nan         nan
int_rate                                9.2080        nan        nan        nan
nan         nan
installment                             8.6440        nan        nan        nan
nan         nan
emp_length                              0.6838        nan        nan        nan
nan         nan
annual_inc                              8.6448        nan        nan        nan
nan         nan
dti                                     7.6954        nan        nan        nan
nan         nan
delinq_2yrs                             0.6567        nan        nan        nan
nan         nan
inq_last_6mths                          0.6121        nan        nan        nan
nan         nan
pub_rec                                -0.7741        nan        nan        nan
nan         nan
total_acc                               8.3401        nan        nan        nan
nan         nan
acc_open_past_24mths                    1.8483        nan        nan        nan
nan         nan
```

```
delinq_amnt                          2.2746       nan        nan        nan
nan         nan
mort_acc                            -0.6124       nan        nan        nan
nan         nan
mths_since_recent_inq                0.5908       nan        nan        nan
nan         nan
num_bc_tl                            1.8676       nan        nan        nan
nan         nan
num_il_tl                           -0.1047       nan        nan        nan
nan         nan
num_op_rev_tl                        6.5538       nan        nan        nan
nan         nan
pub_rec_bankruptcies                 1.5031       nan        nan        nan
nan         nan
tax_liens                            1.0659       nan        nan        nan
nan         nan
total_bc_limit                       8.6655       nan        nan        nan
nan         nan
total_il_high_credit_limit          -0.1629       nan        nan        nan
nan         nan
average_fico                         8.6440       nan        nan        nan
nan         nan
term_ 60 months                      1.5023       nan        nan        nan
nan         nan
sub_grade_A2                        -7.8090       nan        nan        nan
nan         nan
sub_grade_A3                        -7.7137       nan        nan        nan
nan         nan
sub_grade_A4                        -8.5087       nan        nan        nan
nan         nan
sub_grade_A5                        -8.4889       nan        nan        nan
nan         nan
sub_grade_B1                        -8.5570       nan        nan        nan
nan         nan
sub_grade_B2                        -7.9949       nan        nan        nan
nan         nan
sub_grade_B3                        -7.4269       nan        nan        nan
nan         nan
sub_grade_B4                        -6.7166       nan        nan        nan
nan         nan
sub_grade_B5                        -6.1513       nan        nan        nan
nan         nan
sub_grade_C1                        -5.5584       nan        nan        nan
nan         nan
sub_grade_C2                        -5.0456       nan        nan        nan
nan         nan
sub_grade_C3                        -4.6113       nan        nan        nan
nan         nan
sub_grade_C4                        -4.0045       nan        nan        nan
nan         nan
sub_grade_C5                        -3.7417       nan        nan        nan
nan         nan
sub_grade_D1                        -3.3086       nan        nan        nan
nan         nan
sub_grade_D2                        -2.6914       nan        nan        nan
nan         nan
sub_grade_D3                        -2.1063       nan        nan        nan
```

```
nan         nan
sub_grade_D4                          -1.5346       nan        nan        nan
nan         nan
sub_grade_D5                          -0.8582       nan        nan        nan
nan         nan
sub_grade_E1                           0.1803       nan        nan        nan
nan         nan
sub_grade_E2                           0.6826       nan        nan        nan
nan         nan
sub_grade_E3                           1.3696       nan        nan        nan
nan         nan
sub_grade_E4                           1.8183       nan        nan        nan
nan         nan
sub_grade_E5                           2.0509       nan        nan        nan
nan         nan
sub_grade_F1                           2.6286       nan        nan        nan
nan         nan
sub_grade_F2                           3.3233       nan        nan        nan
nan         nan
sub_grade_F3                           3.3968       nan        nan        nan
nan         nan
sub_grade_F4                           3.7302       nan        nan        nan
nan         nan
sub_grade_F5                           3.9062       nan        nan        nan
nan         nan
sub_grade_G1                           3.9118       nan        nan        nan
nan         nan
sub_grade_G2                           4.0471       nan        nan        nan
nan         nan
sub_grade_G3                           4.1598       nan        nan        nan
nan         nan
sub_grade_G4                           4.1936       nan        nan        nan
nan         nan
sub_grade_G5                           4.2442       nan        nan        nan
nan         nan
home_ownership_MORTGAGE              -19.8524       nan        nan        nan
nan         nan
home_ownership_NONE                    4.2886       nan        nan        nan
nan         nan
home_ownership_OTHER                   4.2898       nan        nan        nan
nan         nan
home_ownership_OWN                   -19.4797       nan        nan        nan
nan         nan
home_ownership_RENT                  -18.3758       nan        nan        nan
nan         nan
verification_status_Source Verified   -1.2702       nan        nan        nan
nan         nan
verification_status_Verified          -1.4531       nan        nan        nan
nan         nan
purpose_credit_card                   -7.1392       nan        nan        nan
nan         nan
purpose_debt_consolidation            -6.3703       nan        nan        nan
nan         nan
purpose_educational                    4.3196       nan        nan        nan
nan         nan
purpose_home_improvement              -9.7802       nan        nan        nan
nan         nan
```

```
purpose_house                              0.0230        nan        nan        nan
nan         nan
purpose_major_purchase                    -6.9460        nan        nan        nan
nan         nan
purpose_medical                           -3.5211        nan        nan        nan
nan         nan
purpose_moving                            -0.5347        nan        nan        nan
nan         nan
purpose_other                             -9.8886        nan        nan        nan
nan         nan
purpose_renewable_energy                   3.7960        nan        nan        nan
nan         nan
purpose_small_business                    -1.3697        nan        nan        nan
nan         nan
purpose_vacation                          -1.1498        nan        nan        nan
nan         nan
purpose_wedding                            3.7747        nan        nan        nan
nan         nan
addr_state_AL                              0.4408        nan        nan        nan
nan         nan
addr_state_AR                              1.7599        nan        nan        nan
nan         nan
addr_state_AZ                             -2.6174        nan        nan        nan
nan         nan
addr_state_CA                             -2.9076        nan        nan        nan
nan         nan
addr_state_CO                             -3.2649        nan        nan        nan
nan         nan
addr_state_CT                             -1.4245        nan        nan        nan
nan         nan
addr_state_DC                              2.9073        nan        nan        nan
nan         nan
addr_state_DE                              2.9198        nan        nan        nan
nan         nan
addr_state_FL                             -2.9972        nan        nan        nan
nan         nan
addr_state_GA                             -3.7059        nan        nan        nan
nan         nan
addr_state_HI                              1.8640        nan        nan        nan
nan         nan
addr_state_IA                              4.3208        nan        nan        nan
nan         nan
addr_state_ID                              3.3052        nan        nan        nan
nan         nan
addr_state_IL                             -4.2040        nan        nan        nan
nan         nan
addr_state_IN                             -1.0688        nan        nan        nan
nan         nan
addr_state_KS                              0.6291        nan        nan        nan
nan         nan
addr_state_KY                              0.6212        nan        nan        nan
nan         nan
addr_state_LA                              0.7097        nan        nan        nan
nan         nan
addr_state_MA                             -2.6008        nan        nan        nan
nan         nan
addr_state_MD                             -2.1459        nan        nan        nan
```

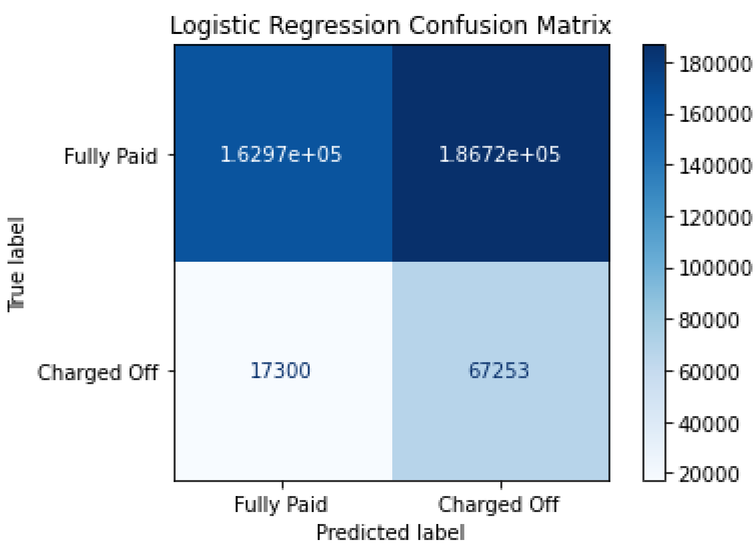
```
nan          nan
addr_state_ME                              3.1254       nan        nan        nan
nan          nan
addr_state_MI                             -2.6974       nan        nan        nan
nan          nan
addr_state_MN                             -1.5419       nan        nan        nan
nan          nan
addr_state_MO                             -0.7965       nan        nan        nan
nan          nan
addr_state_MS                              2.4667       nan        nan        nan
nan          nan
addr_state_MT                              2.8083       nan        nan        nan
nan          nan
addr_state_NC                             -2.6209       nan        nan        nan
nan          nan
addr_state_ND                              3.5380       nan        nan        nan
nan          nan
addr_state_NE                              2.9653       nan        nan        nan
nan          nan
addr_state_NH                              1.6866       nan        nan        nan
nan          nan
addr_state_NJ                             -3.7530       nan        nan        nan
nan          nan
addr_state_NM                              2.0615       nan        nan        nan
nan          nan
addr_state_NV                             -0.8124       nan        nan        nan
nan          nan
addr_state_NY                             -3.0850       nan        nan        nan
nan          nan
addr_state_OH                             -3.4002       nan        nan        nan
nan          nan
addr_state_OK                              1.2050       nan        nan        nan
nan          nan
addr_state_OR                             -1.3092       nan        nan        nan
nan          nan
addr_state_PA                             -3.2496       nan        nan        nan
nan          nan
addr_state_RI                              2.0145       nan        nan        nan
nan          nan
addr_state_SC                             -0.8578       nan        nan        nan
nan          nan
addr_state_SD                              3.3753       nan        nan        nan
nan          nan
addr_state_TN                             -1.1177       nan        nan        nan
nan          nan
addr_state_TX                             -2.8919       nan        nan        nan
nan          nan
addr_state_UT                              0.9898       nan        nan        nan
nan          nan
addr_state_VA                             -2.8245       nan        nan        nan
nan          nan
addr_state_VT                              3.0787       nan        nan        nan
nan          nan
addr_state_WA                             -3.4040       nan        nan        nan
nan          nan
addr_state_WI                             -0.6818       nan        nan        nan
nan          nan
```

```
addr_state_WV                            2.3983        nan         nan         nan
nan         nan
addr_state_WY                            3.2252        nan         nan         nan
nan         nan
application_type_Joint App               1.2391        nan         nan         nan
nan         nan
================================================================================
========================
```

6 Scaled Smote

In [121]:
```
logreg = LogisticRegression(C=1e12,fit_intercept=False,solver='liblinear',penalty
log_model = logreg.fit(x_train_smscaled, y_smote)
y_hat_log = logreg.predict(x_test_scaled)
y_score_log = log_model.decision_function(x_test_scaled)
fpr,tpr,thresholds = roc_curve(y_test,y_score_log)

print('AUC: {}'.format(auc(fpr, tpr)))
cf = confusion_matrix(y_test,y_hat_log)

plot_confusion_matrix(log_model,x_test_scaled,y_test,cmap=plt.cm.Blues,
                     display_labels=["Fully Paid", "Charged Off"],
                      values_format=".5g")
plt.title("Logistic Regression Confusion Matrix")

print(confusion_matrix(y_test, y_hat_log))
print(classification_report(y_test, y_hat_log))
```
executed in 44m 27s, finished 03:07:15 2021-04-07

```
C:\Users\sergi\anaconda3\envs\learn-env\lib\site-packages\sklearn\svm\_base.py:
976: ConvergenceWarning: Liblinear failed to converge, increase the number of i
terations.
  warnings.warn("Liblinear failed to converge, increase "

AUC: 0.6986879801984937
[[162967 186715]
 [ 17300  67253]]
              precision    recall  f1-score   support

           0       0.90      0.47      0.62    349682
           1       0.26      0.80      0.40     84553

    accuracy                           0.53    434235
   macro avg       0.58      0.63      0.51    434235
weighted avg       0.78      0.53      0.57    434235
```



In [ ]:
```
log
```

In [127]:

```
logit_model=sm.Logit(y_smote,x_train_smscaled)
result=logit_model.fit(method='lbfgs',maxiter=300)
print(result.summary())
```

executed in 3m 22s, finished 03:51:08 2021-04-07

```
C:\Users\sergi\anaconda3\envs\learn-env\lib\site-packages\statsmodels\discrete
\discrete_model.py:1799: RuntimeWarning: overflow encountered in exp
  return 1/(1+np.exp(-X))
C:\Users\sergi\anaconda3\envs\learn-env\lib\site-packages\statsmodels\base\mode
l.py:547: HessianInversionWarning: Inverting hessian failed, no bse or cov_para
ms available
  warnings.warn('Inverting hessian failed, no bse or cov_params '


                           Logit Regression Results
==============================================================================
Dep. Variable:            loan_status   No. Observations:              2100320
Model:                          Logit   Df Residuals:                  2100192
Method:                           MLE   Df Model:                          127
Date:                Wed, 07 Apr 2021   Pseudo R-squ.:                  0.5840
Time:                        03:51:07   Log-Likelihood:            -6.0558e+05
converged:                       True   LL-Null:                   -1.4558e+06
Covariance Type:            nonrobust   LLR p-value:                     0.000
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
x1             0.3605        nan        nan        nan         nan         nan
x2             0.8966        nan        nan        nan         nan         nan
x3            -0.1496        nan        nan        nan         nan         nan
x4            -0.0293        nan        nan        nan         nan         nan
x5            -0.0792        nan        nan        nan         nan         nan
x6             0.1377        nan        nan        nan         nan         nan
x7             0.0344        nan        nan        nan         nan         nan
x8             0.0362        nan        nan        nan         nan         nan
x9             0.0044        nan        nan        nan         nan         nan
x10           -0.2171        nan        nan        nan         nan         nan
x11            0.1765        nan        nan        nan         nan         nan
x12            0.0041        nan        nan        nan         nan         nan
x13           -0.0779        nan        nan        nan         nan         nan
x14           -0.0458        nan        nan        nan         nan         nan
x15            0.0487        nan        nan        nan         nan         nan
x16            0.0684        nan        nan        nan         nan         nan
x17            0.1111        nan        nan        nan         nan         nan
x18            0.0056        nan        nan        nan         nan         nan
x19            0.0045        nan        nan        nan         nan         nan
x20           -0.1555        nan        nan        nan         nan         nan
x21           -0.0556        nan        nan        nan         nan         nan
x22           -0.1744        nan        nan        nan         nan         nan
x23            0.2135        nan        nan        nan         nan         nan
x24           -0.2225        nan        nan        nan         nan         nan
x25           -0.2305        nan        nan        nan         nan         nan
x26           -0.2581        nan        nan        nan         nan         nan
x27           -0.2812        nan        nan        nan         nan         nan
x28           -0.3113        nan        nan        nan         nan         nan
x29           -0.3324        nan        nan        nan         nan         nan
x30           -0.3545        nan        nan        nan         nan         nan
```

x31	-0.3739	nan	nan	nan	nan	nan
x32	-0.3810	nan	nan	nan	nan	nan
x33	-0.3982	nan	nan	nan	nan	nan
x34	-0.3927	nan	nan	nan	nan	nan
x35	-0.3995	nan	nan	nan	nan	nan
x36	-0.4113	nan	nan	nan	nan	nan
x37	-0.4188	nan	nan	nan	nan	nan
x38	-0.3773	nan	nan	nan	nan	nan
x39	-0.3776	nan	nan	nan	nan	nan
x40	-0.3629	nan	nan	nan	nan	nan
x41	-0.3536	nan	nan	nan	nan	nan
x42	-0.3460	nan	nan	nan	nan	nan
x43	-0.2924	nan	nan	nan	nan	nan
x44	-0.2865	nan	nan	nan	nan	nan
x45	-0.2754	nan	nan	nan	nan	nan
x46	-0.2672	nan	nan	nan	nan	nan
x47	-0.2773	nan	nan	nan	nan	nan
x48	-0.2286	nan	nan	nan	nan	nan
x49	-0.1885	nan	nan	nan	nan	nan
x50	-0.1830	nan	nan	nan	nan	nan
x51	-0.1624	nan	nan	nan	nan	nan
x52	-0.1526	nan	nan	nan	nan	nan
x53	-0.1421	nan	nan	nan	nan	nan
x54	-0.1157	nan	nan	nan	nan	nan
x55	-0.1020	nan	nan	nan	nan	nan
x56	-0.0923	nan	nan	nan	nan	nan
x57	-0.0883	nan	nan	nan	nan	nan
x58	-1.6203	nan	nan	nan	nan	nan
x59	-0.0164	nan	nan	nan	nan	nan
x60	-0.0153	nan	nan	nan	nan	nan
x61	-0.8351	nan	nan	nan	nan	nan
x62	-1.4176	nan	nan	nan	nan	nan
x63	-0.0010	nan	nan	nan	nan	nan
x64	0.0054	nan	nan	nan	nan	nan
x65	-0.4458	nan	nan	nan	nan	nan
x66	-0.5752	nan	nan	nan	nan	nan
x67	-1254.5302	nan	nan	nan	nan	nan
x68	-0.2290	nan	nan	nan	nan	nan
x69	-0.0712	nan	nan	nan	nan	nan
x70	-0.1286	nan	nan	nan	nan	nan
x71	-0.0847	nan	nan	nan	nan	nan
x72	-0.0670	nan	nan	nan	nan	nan
x73	-0.2156	nan	nan	nan	nan	nan
x74	-0.0226	nan	nan	nan	nan	nan
x75	-0.0571	nan	nan	nan	nan	nan
x76	-0.0747	nan	nan	nan	nan	nan
x77	-0.0313	nan	nan	nan	nan	nan
x78	-0.3446	nan	nan	nan	nan	nan
x79	-0.2668	nan	nan	nan	nan	nan
x80	-0.5188	nan	nan	nan	nan	nan
x81	-1.2288	nan	nan	nan	nan	nan
x82	-0.5235	nan	nan	nan	nan	nan
x83	-0.4123	nan	nan	nan	nan	nan
x84	-0.1745	nan	nan	nan	nan	nan
x85	-0.1780	nan	nan	nan	nan	nan
x86	-0.8631	nan	nan	nan	nan	nan
x87	-0.6046	nan	nan	nan	nan	nan

```
x88          -0.2350       nan       nan       nan       nan       nan
x89        -887.0893       nan       nan       nan       nan       nan
x90          -0.1546       nan       nan       nan       nan       nan
x91          -0.6608       nan       nan       nan       nan       nan
x92          -0.4229       nan       nan       nan       nan       nan
x93          -0.3109       nan       nan       nan       nan       nan
x94          -0.3189       nan       nan       nan       nan       nan
x95          -0.3324       nan       nan       nan       nan       nan
x96          -0.5022       nan       nan       nan       nan       nan
x97          -0.4919       nan       nan       nan       nan       nan
x98          -0.1649       nan       nan       nan       nan       nan
x99          -0.5322       nan       nan       nan       nan       nan
x100         -0.4422       nan       nan       nan       nan       nan
x101         -0.4068       nan       nan       nan       nan       nan
x102         -0.2235       nan       nan       nan       nan       nan
x103         -0.1854       nan       nan       nan       nan       nan
x104         -0.5463       nan       nan       nan       nan       nan
x105         -0.1346       nan       nan       nan       nan       nan
x106         -0.1816       nan       nan       nan       nan       nan
x107         -0.2493       nan       nan       nan       nan       nan
x108         -0.6191       nan       nan       nan       nan       nan
x109         -0.2373       nan       nan       nan       nan       nan
x110         -0.3978       nan       nan       nan       nan       nan
x111         -0.9114       nan       nan       nan       nan       nan
x112         -0.5978       nan       nan       nan       nan       nan
x113         -0.2983       nan       nan       nan       nan       nan
x114         -0.3969       nan       nan       nan       nan       nan
x115         -0.5976       nan       nan       nan       nan       nan
x116         -0.2267       nan       nan       nan       nan       nan
x117         -0.3887       nan       nan       nan       nan       nan
x118         -0.1465       nan       nan       nan       nan       nan
x119         -0.4189       nan       nan       nan       nan       nan
x120         -0.9377       nan       nan       nan       nan       nan
x121         -0.2929       nan       nan       nan       nan       nan
x122         -0.5450       nan       nan       nan       nan       nan
x123         -0.1679       nan       nan       nan       nan       nan
x124         -0.5204       nan       nan       nan       nan       nan
x125         -0.3831       nan       nan       nan       nan       nan
x126         -0.2080       nan       nan       nan       nan       nan
x127         -0.1591       nan       nan       nan       nan       nan
x128         -0.0057       nan       nan       nan       nan       nan
==============================================================================

C:\Users\sergi\anaconda3\envs\learn-env\lib\site-packages\statsmodels\discrete
\discrete_model.py:1799: RuntimeWarning: overflow encountered in exp
  return 1/(1+np.exp(-X))
```

In [126]:
```
pd.DataFrame(x_train_smscaled,columns=x_train.columns).describe()
```
executed in 12.7s, finished 03:21:41 2021-04-07

Out[126]:

	loan_amnt	int_rate	installment	emp_length	annual_inc	dti	deli
count	2.100320e+06	2.100320e+06	2.100320e+06	2.100320e+06	2.100320e+06	2.100320e+06	2.1
mean	1.013553e-16	6.365493e-17	-3.274765e-18	-4.671276e-17	-2.648906e-17	-2.509119e-16	2.3
std	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.0
min	-1.576928e+00	-1.845820e+00	-1.700920e+00	-1.430622e+00	-5.469043e-01	-1.561032e+00	-3.8
25%	-7.998116e-01	-6.713386e-01	-7.250330e-01	-8.386436e-01	-2.221268e-01	-5.102466e-01	-3.8
50%	-2.075468e-01	-9.248727e-02	-2.462688e-01	4.932387e-02	-7.778118e-02	-5.098657e-02	-3.8
75%	5.323876e-01	6.038122e-01	5.564958e-01	1.233281e+00	1.026508e-01	4.421198e-01	-3.2
max	2.752720e+00	3.540015e+00	4.764545e+00	1.233281e+00	7.933537e+02	7.512994e+01	4.6

8 rows × 128 columns

In [129]:
```
pd.DataFrame(x_train_scaled,columns=x_train.columns).describe()
```
executed in 9.09s, finished 04:05:45 2021-04-07

Out[129]:

	loan_amnt	int_rate	installment	emp_length	annual_inc	dti	d
count	1.302702e+06	1.302702e+06	1.302702e+06	1.302702e+06	1.302702e+06	1.302702e+06	
mean	3.830063e-17	1.374985e-16	-3.542618e-17	-2.438652e-17	-4.310355e-17	-1.345049e-16	
std	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	
min	-1.525434e+00	-1.622500e+00	-1.641543e+00	-1.409350e+00	-5.925288e-01	-1.491245e+00	-
25%	-7.465492e-01	-7.588892e-01	-7.247475e-01	-1.129870e+00	-2.358661e-01	-5.120570e-01	-
50%	-3.014723e-01	-9.775183e-02	-2.520840e-01	-1.195406e-02	-9.837554e-02	-6.518439e-02	-
75%	5.886815e-01	5.840460e-01	5.462837e-01	1.105962e+00	1.144905e-01	4.353733e-01	-
max	2.814066e+00	3.683127e+00	4.771836e+00	1.105962e+00	8.356668e+02	7.412171e+01	-

8 rows × 128 columns

In [134]:
```
y_train.value_counts()
```
executed in 40ms, finished 04:14:17 2021-04-07

Out[134]:
```
0    1050160
1     252542
Name: loan_status, dtype: int64
```

In [ ]:

In [ ]:

In [ ]: