Lending Club Loan Data Project

SFDAT28

Eugene Berson

Summary

Goal: To be able to predict if a Lending Club loan will default or be paid in full

Process:

1. EDA
2. Feature Selection
3. Experiment with multiple models
    a. Logistic Regression
    b. KNN
    c. NLP
    d. Others to come

EDA: Removed Columns with NaN values

inq_last_12m	Number of credit inquiries in past 12 months
total_bal_il	Total current balance of all installment accounts
dti_joint	A ratio calculated using the co-borrowers' total monthly payments on the total debt obligations, excluding mortgages and the requested LC loan, divided by the co-borrowers' combined self-reported monthly income
verified_status_joint	Indicates if the co-borrowers' joint income was verified by LC, not verified, or if the income source was verified
total_cu_tl	Number of finance trades
open_acc_6m	Number of open trades in last 6 months
open_il_6m	Number of currently active installment trades
open_il_12m	Number of installment accounts opened in past 12 months
open_il_24m	Number of installment accounts opened in past 24 months
mths_since_rcnt_il	Months since most recent installment accounts opened
il_util	Ratio of total current balance to high credit/credit limit on all install acct
open_rv_12m	Number of revolving trades opened in past 12 months
open_rv_24m	Number of revolving trades opened in past 24 months
max_bal_bc	Maximum current balance owed on all revolving accounts
all_util	Balance to credit limit on all trades
inq_fi	Number of personal finance inquiries
annual_inc_joint	The combined self-reported annual income provided by the co-borrowers during registration

EDA: change categorical values from string to number

Loan Status	
Fully Paid	0
Charged Off	1
Default	1
In Grace Period	2
Issued	2
Does not meet the credit policy. Status: Charged Off	2
Current	2
Does not meet the credit policy. Status: Fully Paid	2
Late (31-120 days)	2
Late (16-30 days)	2

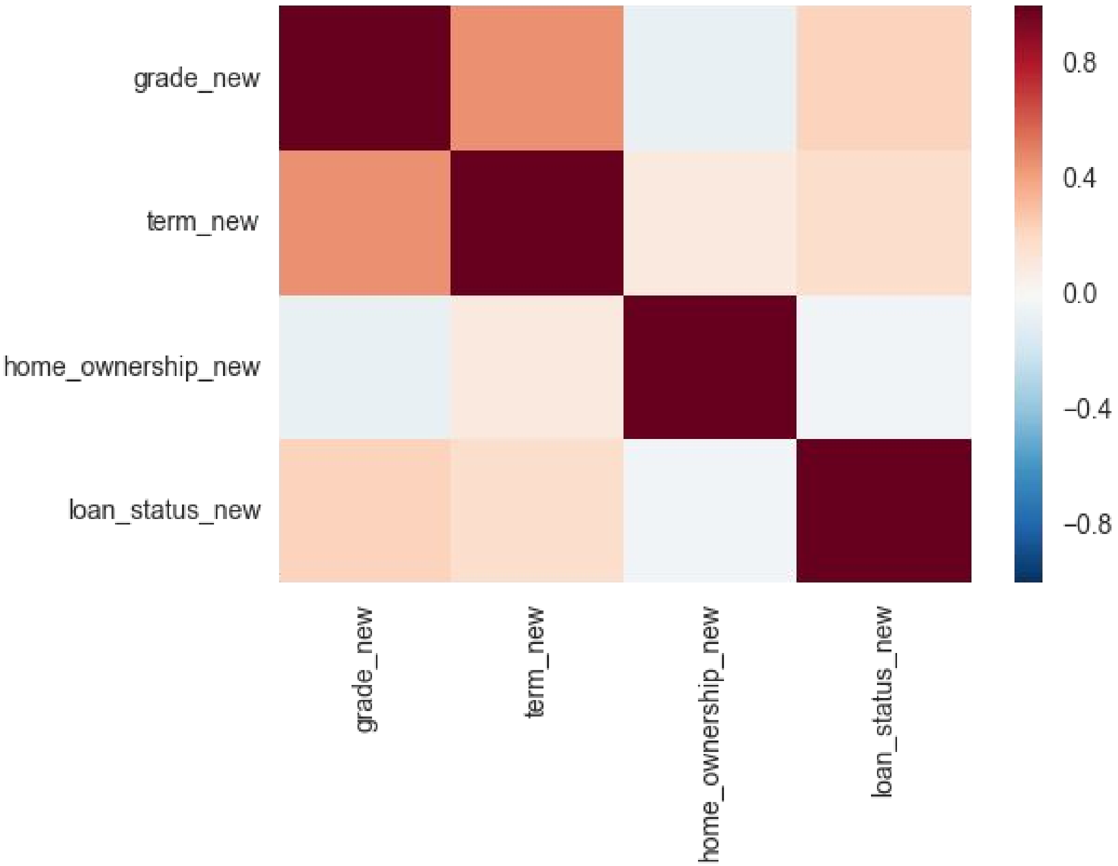
Grade	
A	1
B	2
C	3
D	4
E	5
F	6
G	7

Home_Ownership	
RENT	0
OWN	1
MORTGAGE	1
OTHER	2
NONE	2
ANY	2

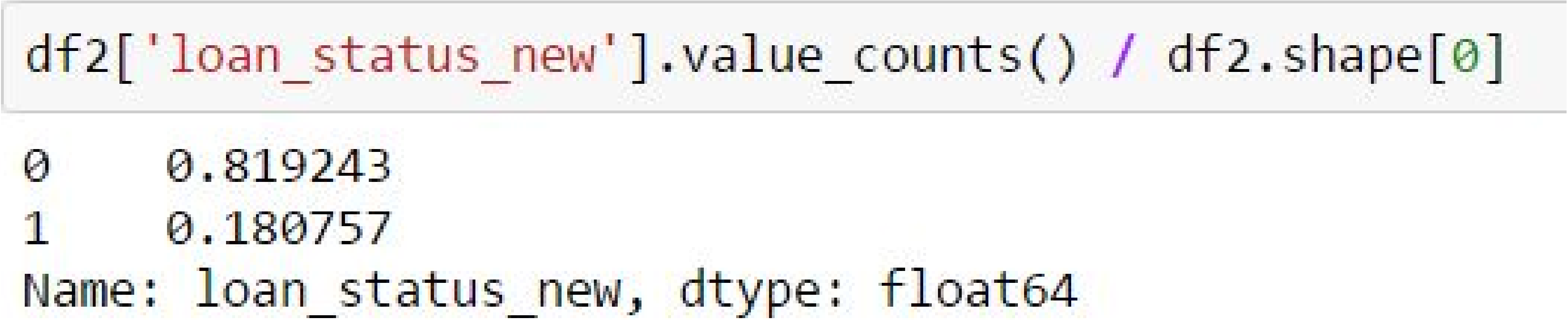
Term	
36 months	3
60 months	5

Feature Selection

- Grade and Term have a moderate correlation with Loan Status



Null Value

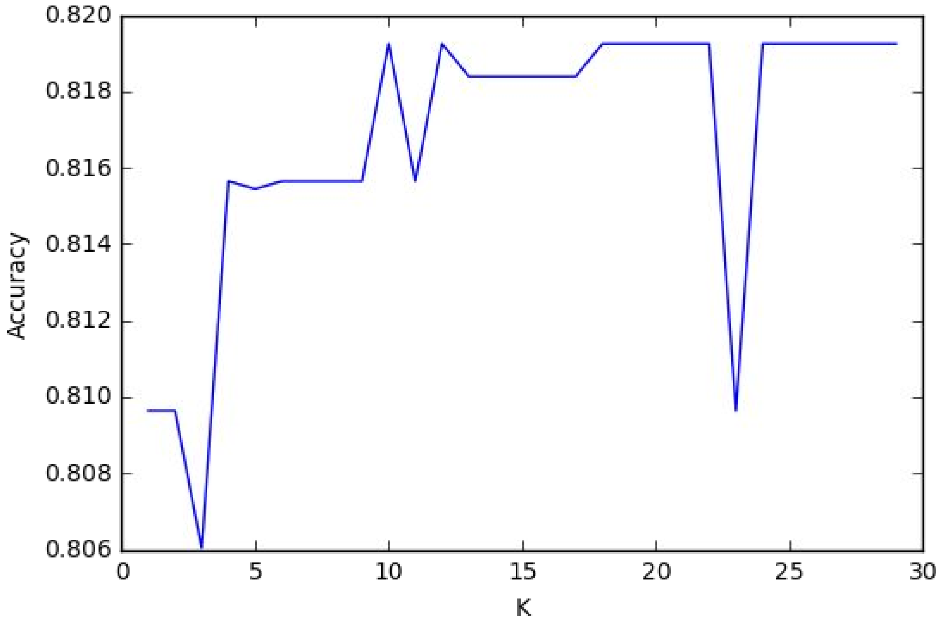


```
df2['loan_status_new'].value_counts() / df2.shape[0]
```

```
0    0.819243
1    0.180757
Name: loan_status_new, dtype: float64
```

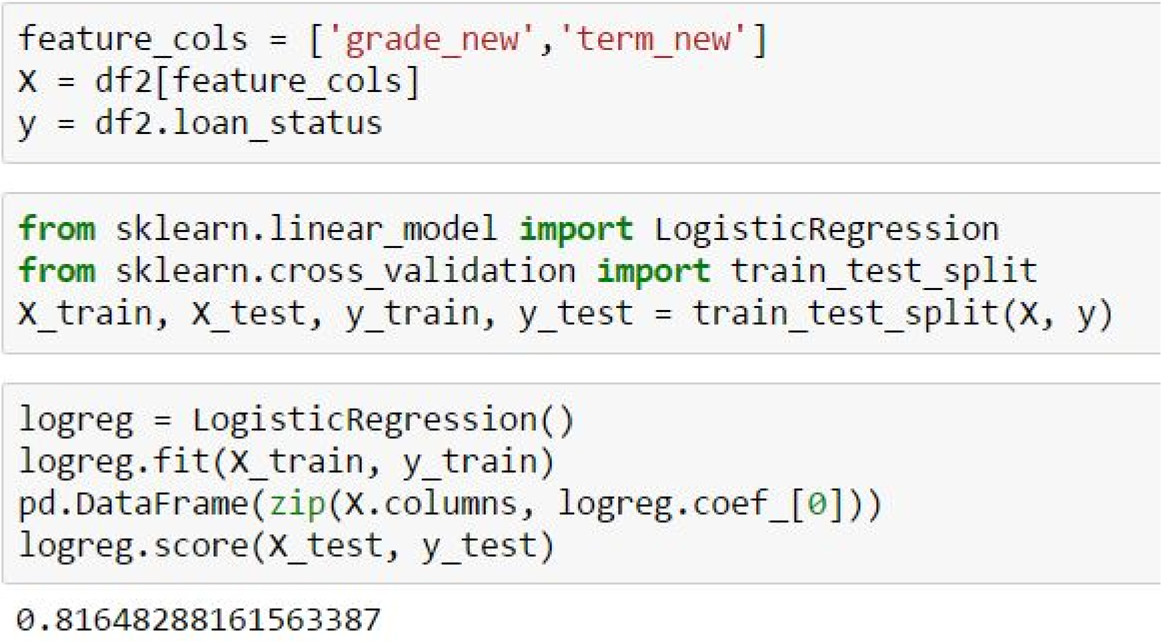
KNN

- Does not beat Null Value of 81.9243%



Logistic Regression

- Does not beat Null Value of 81.9243%



```
feature_cols = ['grade_new','term_new']
X = df2[feature_cols]
y = df2.loan_status
```

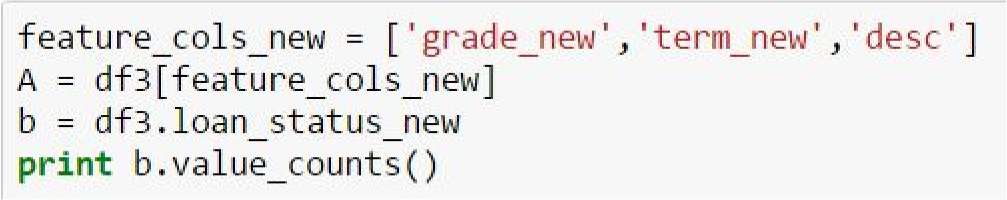
```
from sklearn.linear_model import LogisticRegression
from sklearn.cross_validation import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y)
```

```
logreg = LogisticRegression()
logreg.fit(X_train, y_train)
pd.DataFrame(zip(X.columns, logreg.coef_[0]))
logreg.score(X_test, y_test)
```

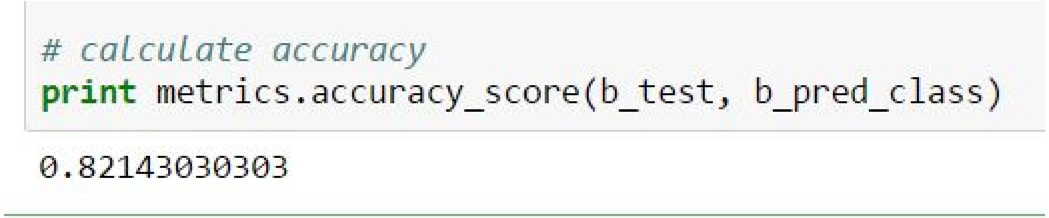
```
0.8164828816156338
```

NLP

By analyzing the loan description, we are actually able to get a better prediction of the loan status than by looking at other factors



```
feature_cols_new = ['grade_new','term_new','desc']
A = df3[feature_cols_new]
b = df3.loan_status_new
print b.value_counts()
```



```
# calculate accuracy
print metrics.accuracy_score(b_test, b_pred_class)
```

```
0.82143030303
```

Resources

https://www.kaggle.com/wendykan/lending-club-loan-data