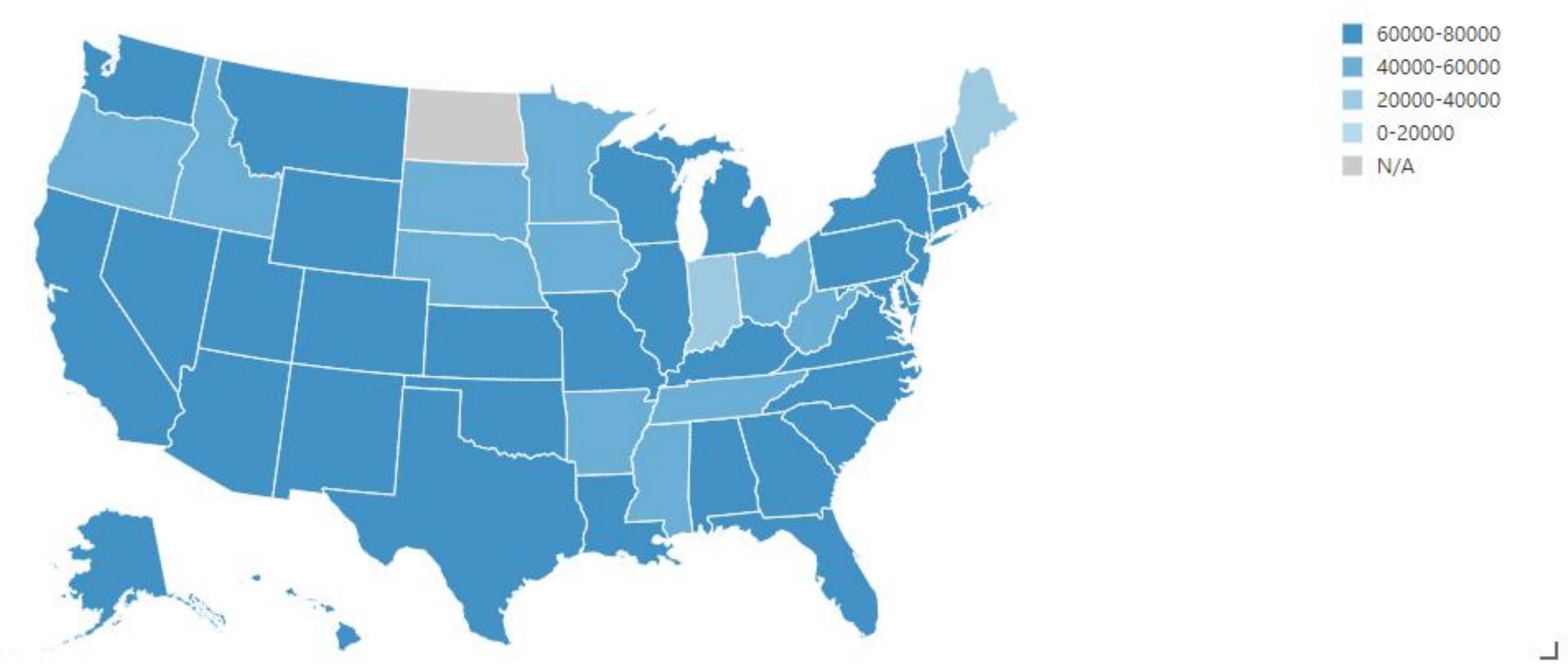
Data Science

Evaluating Risk for Loan Approvals on LendingClub



The self-reported annual income provided by the borrower during registration.

The number of open credit lines in the borrower's credit file.

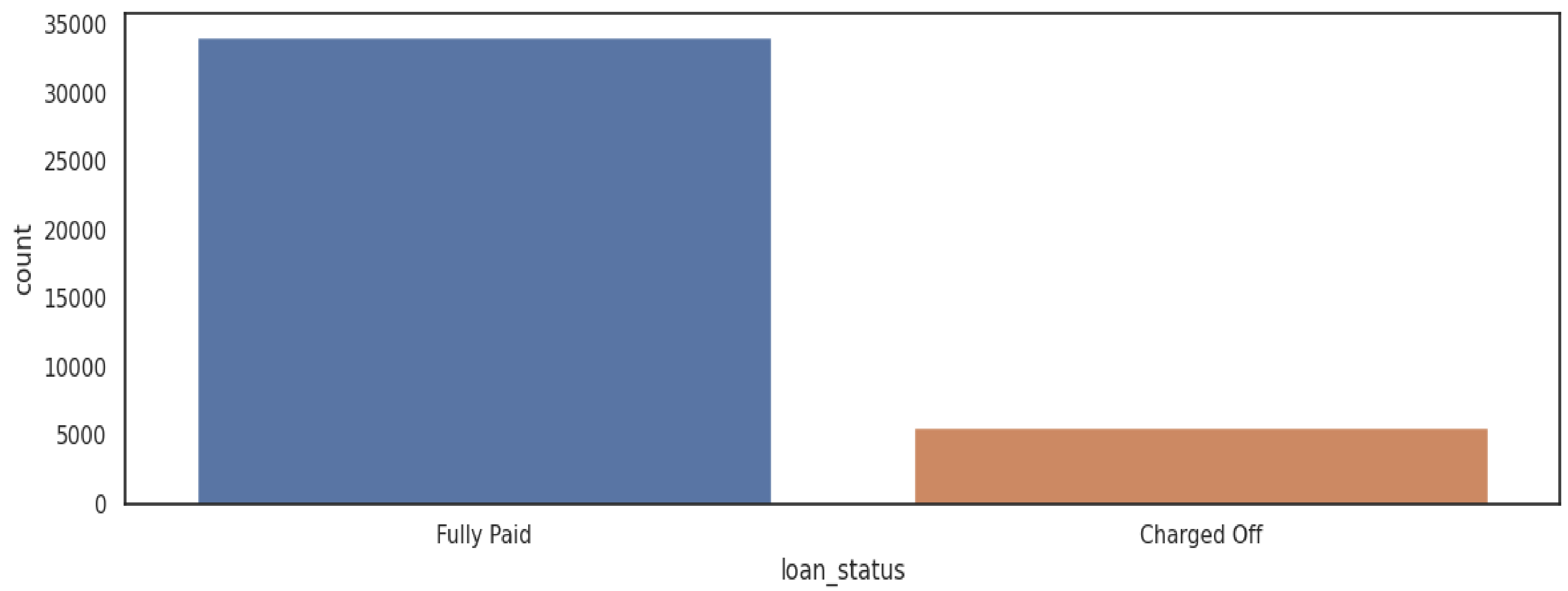
The total number of credit lines currently in the borrower's credit file

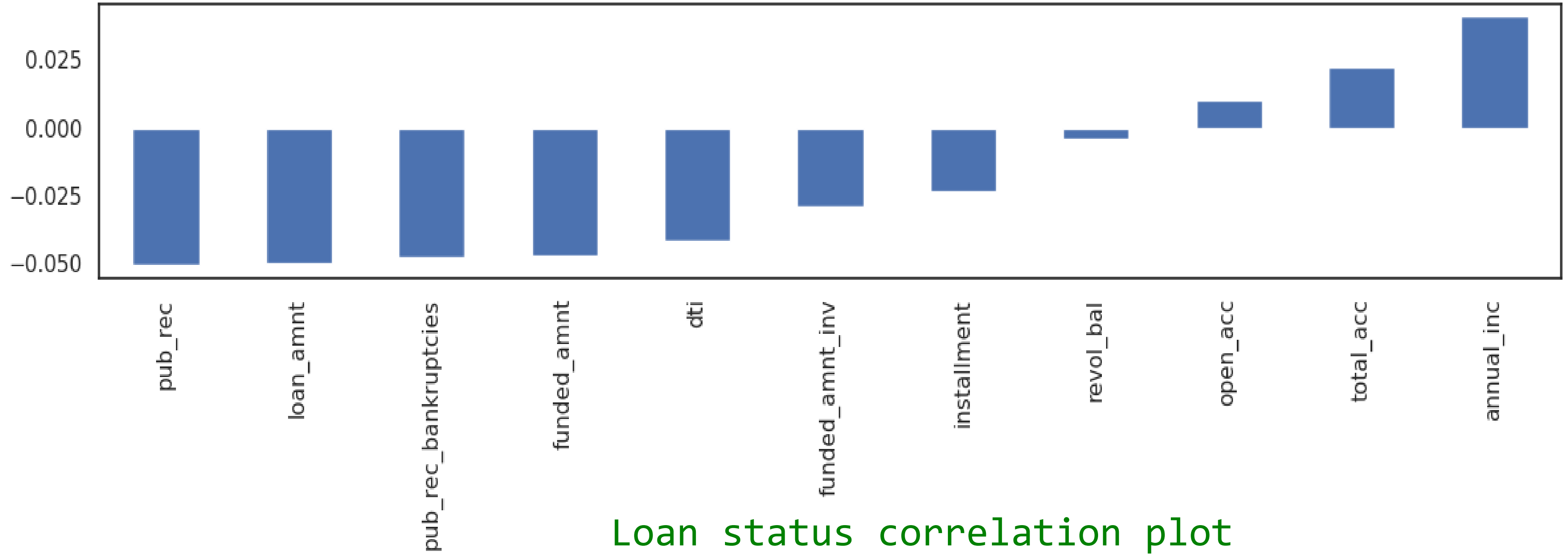
Total credit revolving balance

Debt-to-income ratio

Loan amount granted

Number of derogatory public records

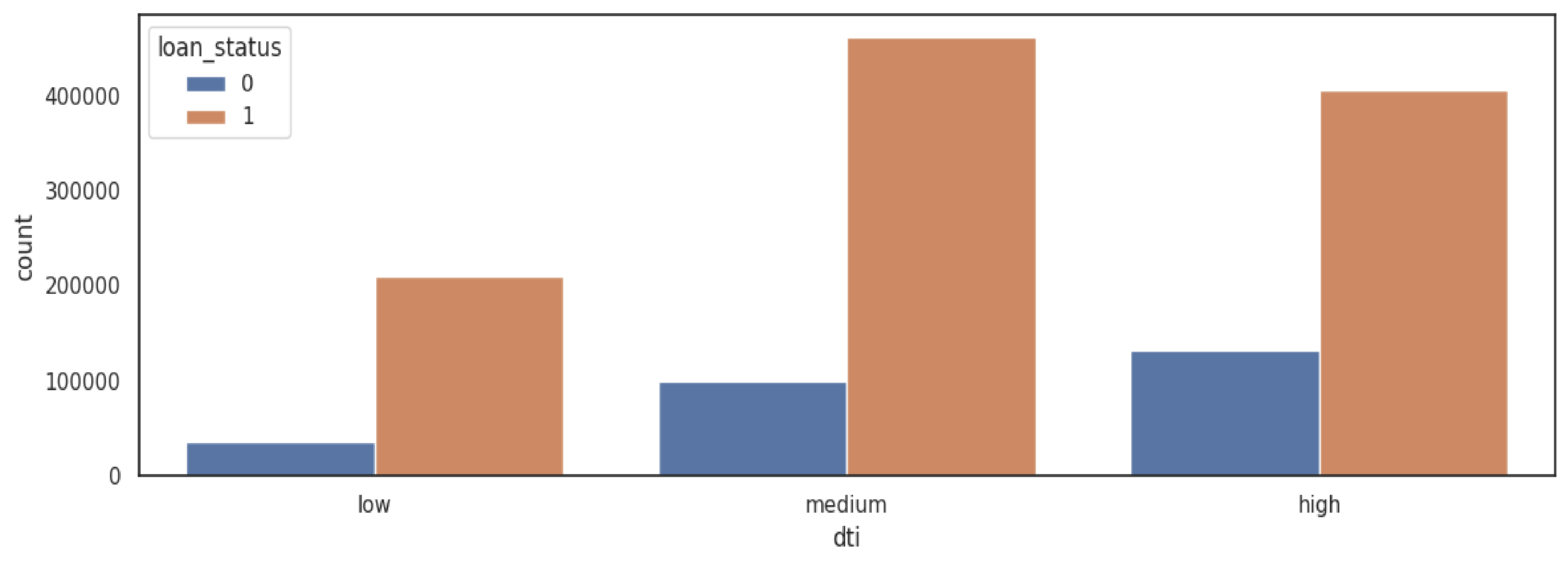




Loan status correlation plot

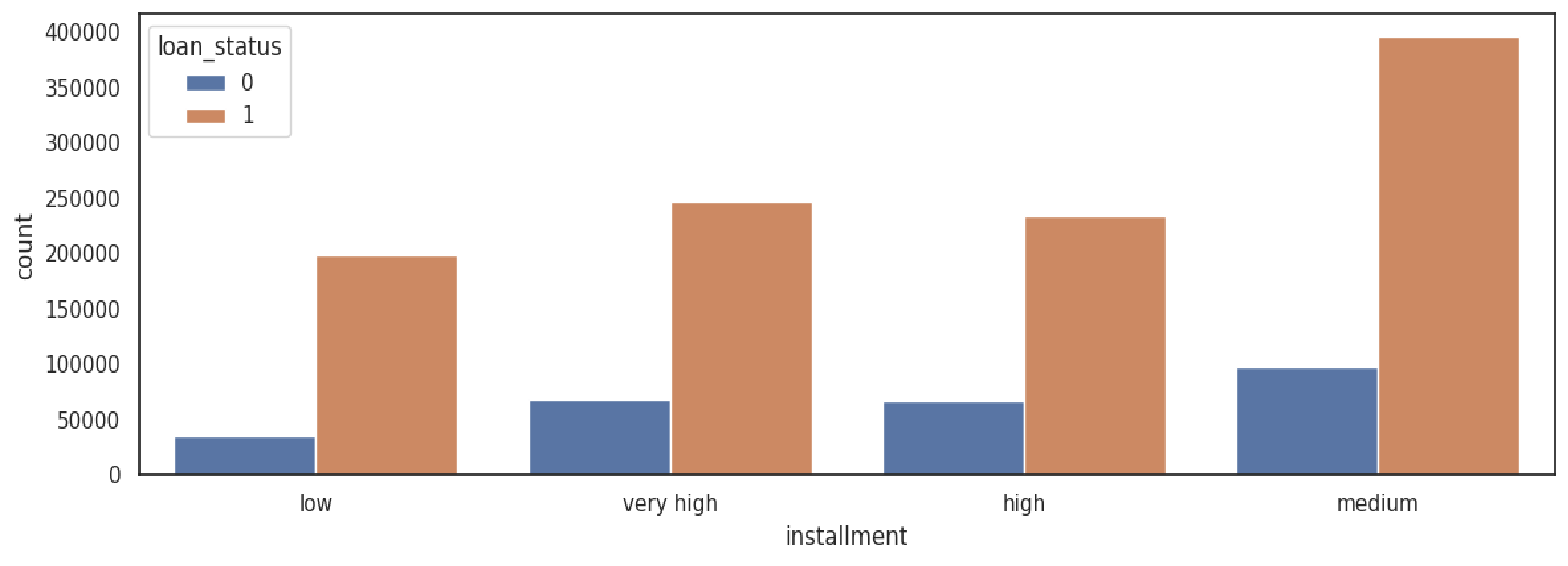
High DTI translates into higher default rates

A ratio calculated using the borrower's total monthly debt payments



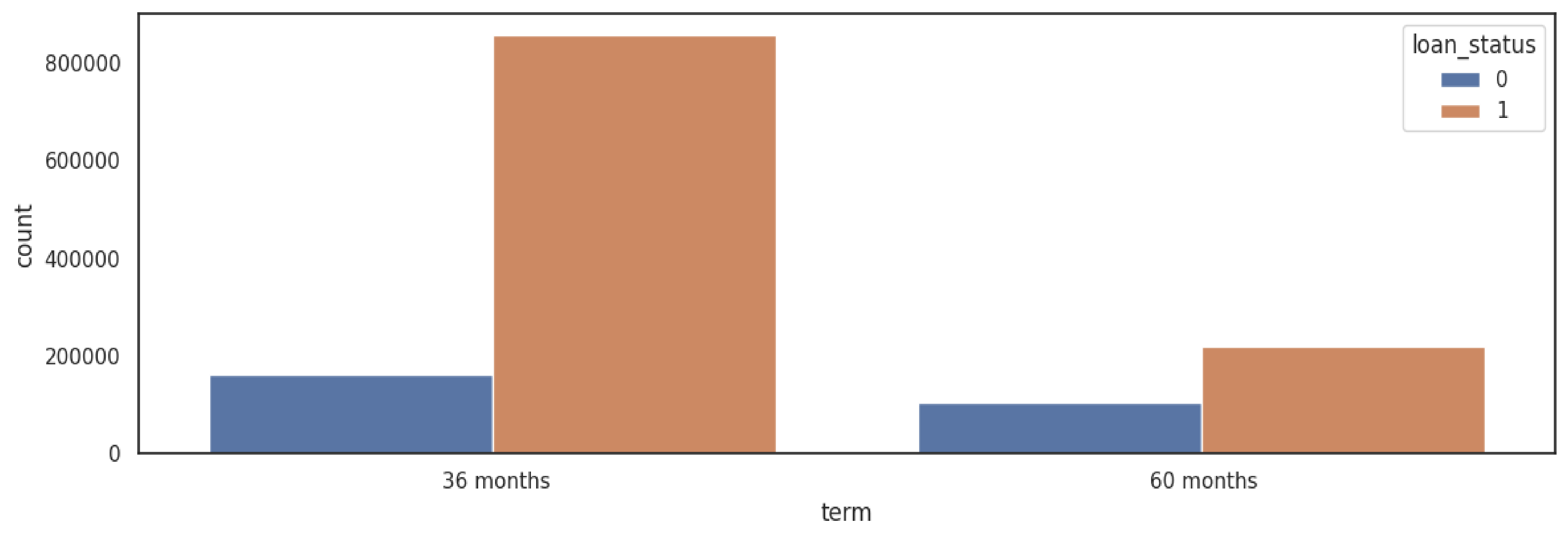
Higher the installment amount, higher the default rate

The monthly payment owed by the borrower if the loan originates.



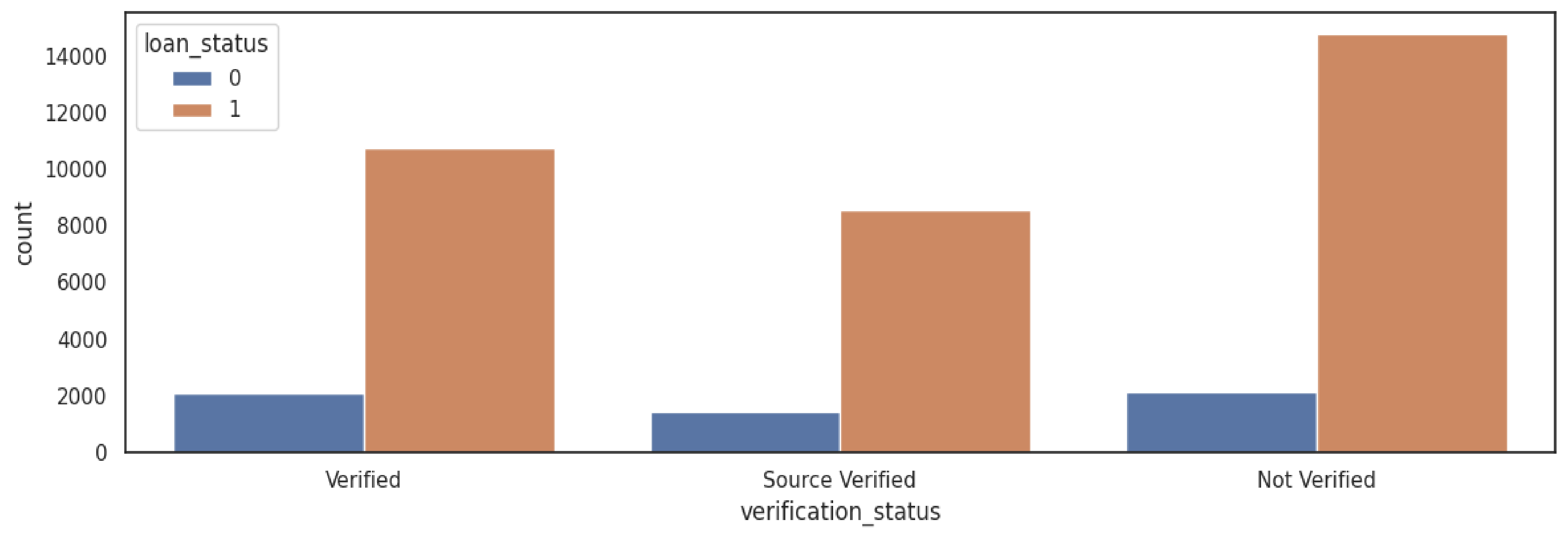
60 months term loan are defaulted more often

The number of payments on the loan. Values are in months and can be either 36 or 60.



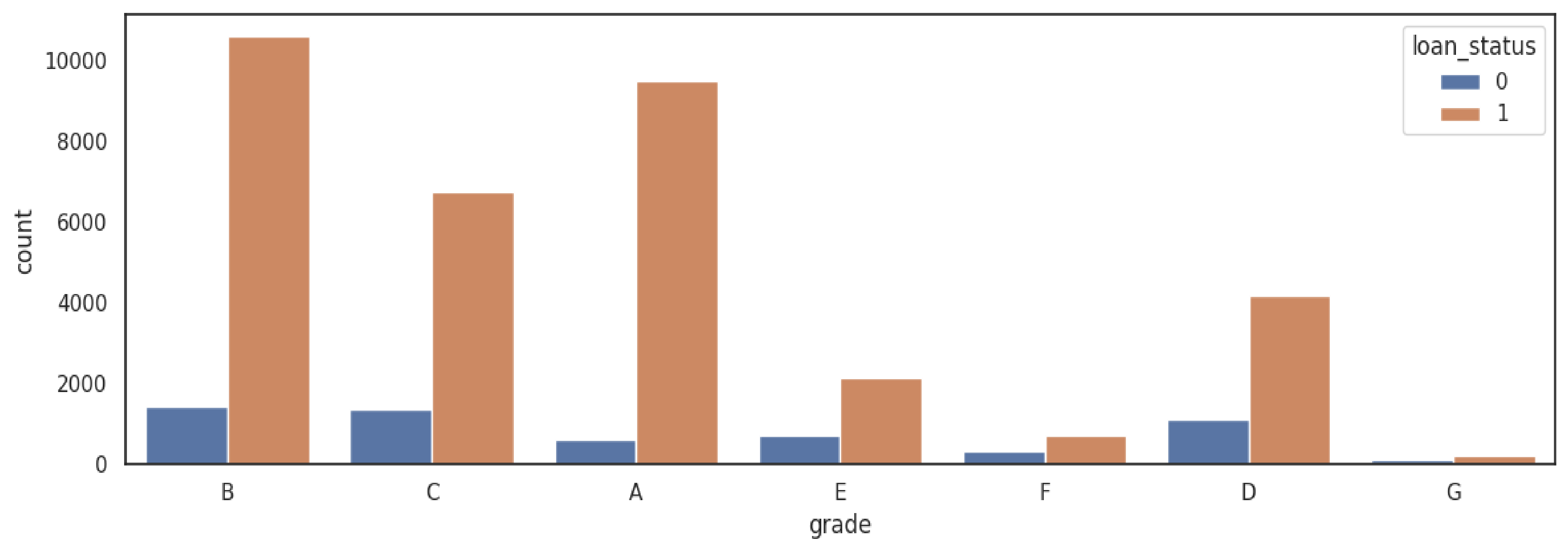
Verified accounts defaulted in large number

Indicates if income was verified by LC, not verified, or if the income source was verified



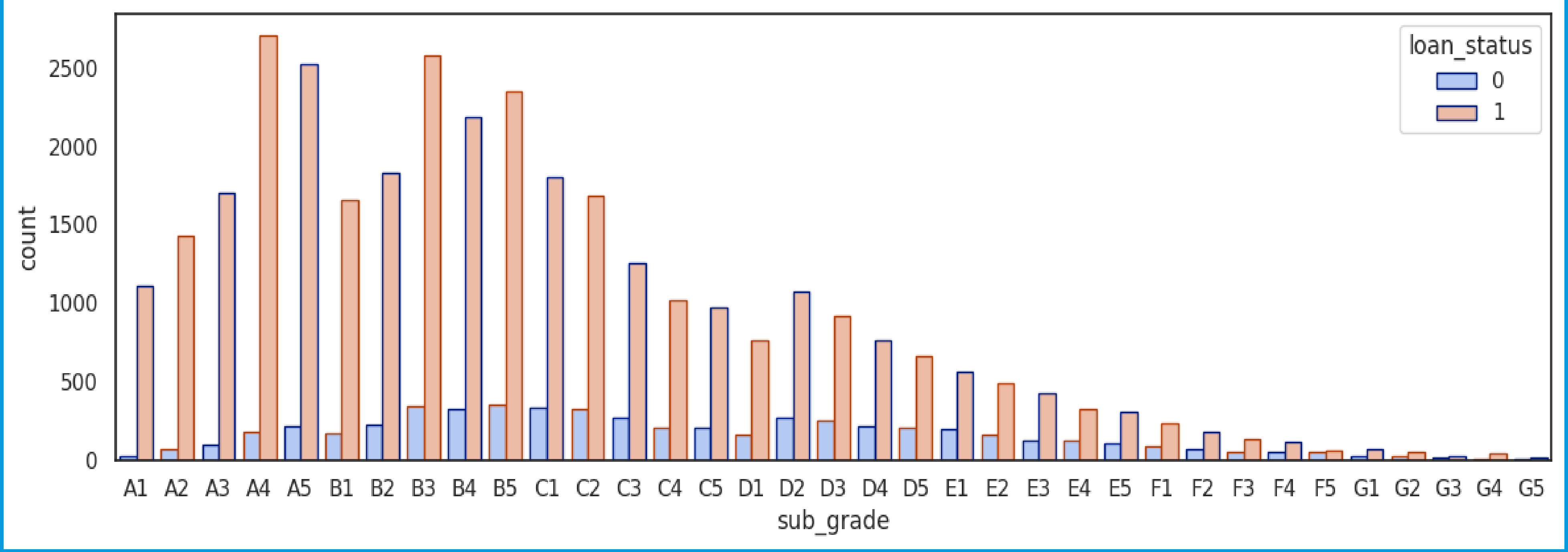
LC assigned loan grade

The 'E' and 'F' grade loans are defaulted more often.

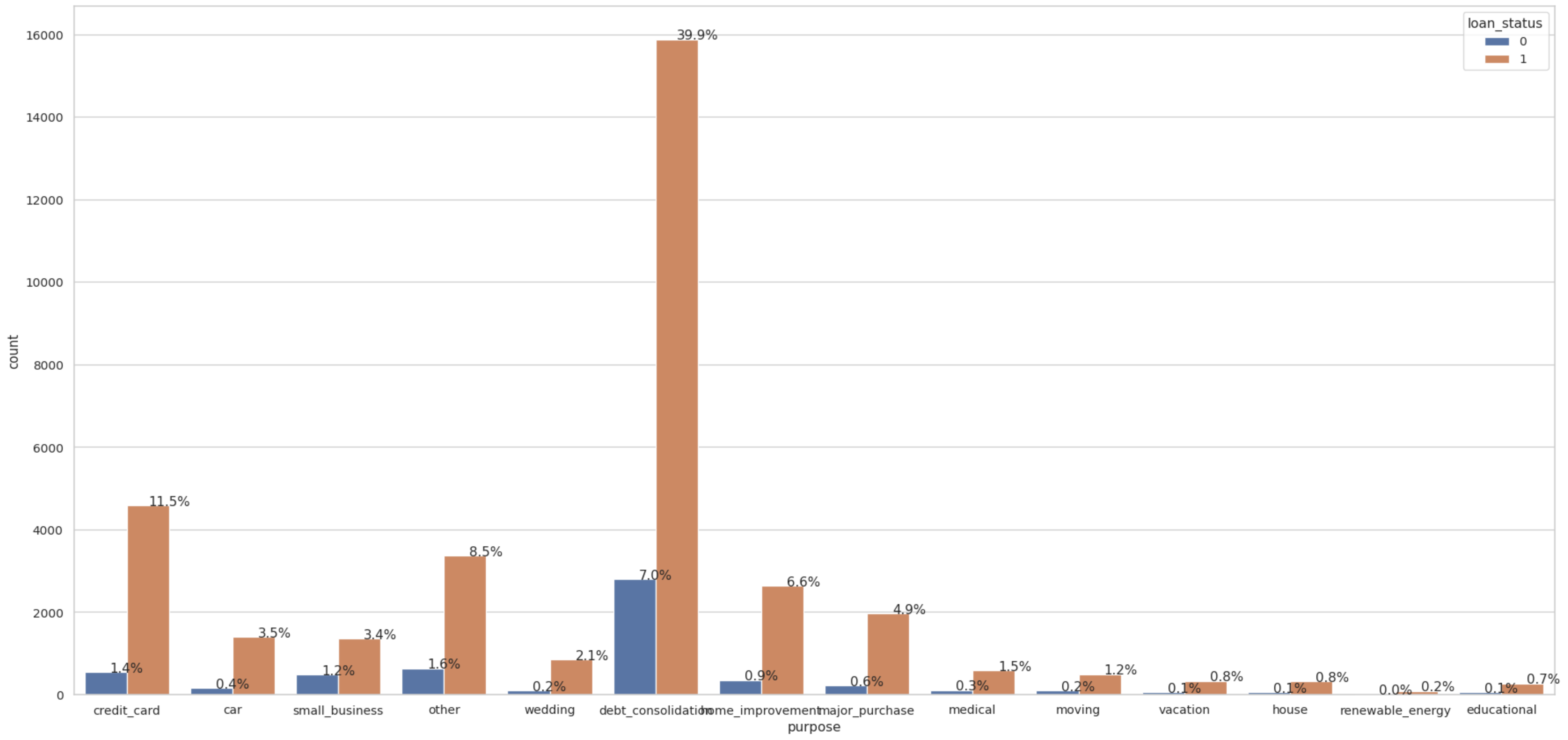


LC assigned loan subgrade

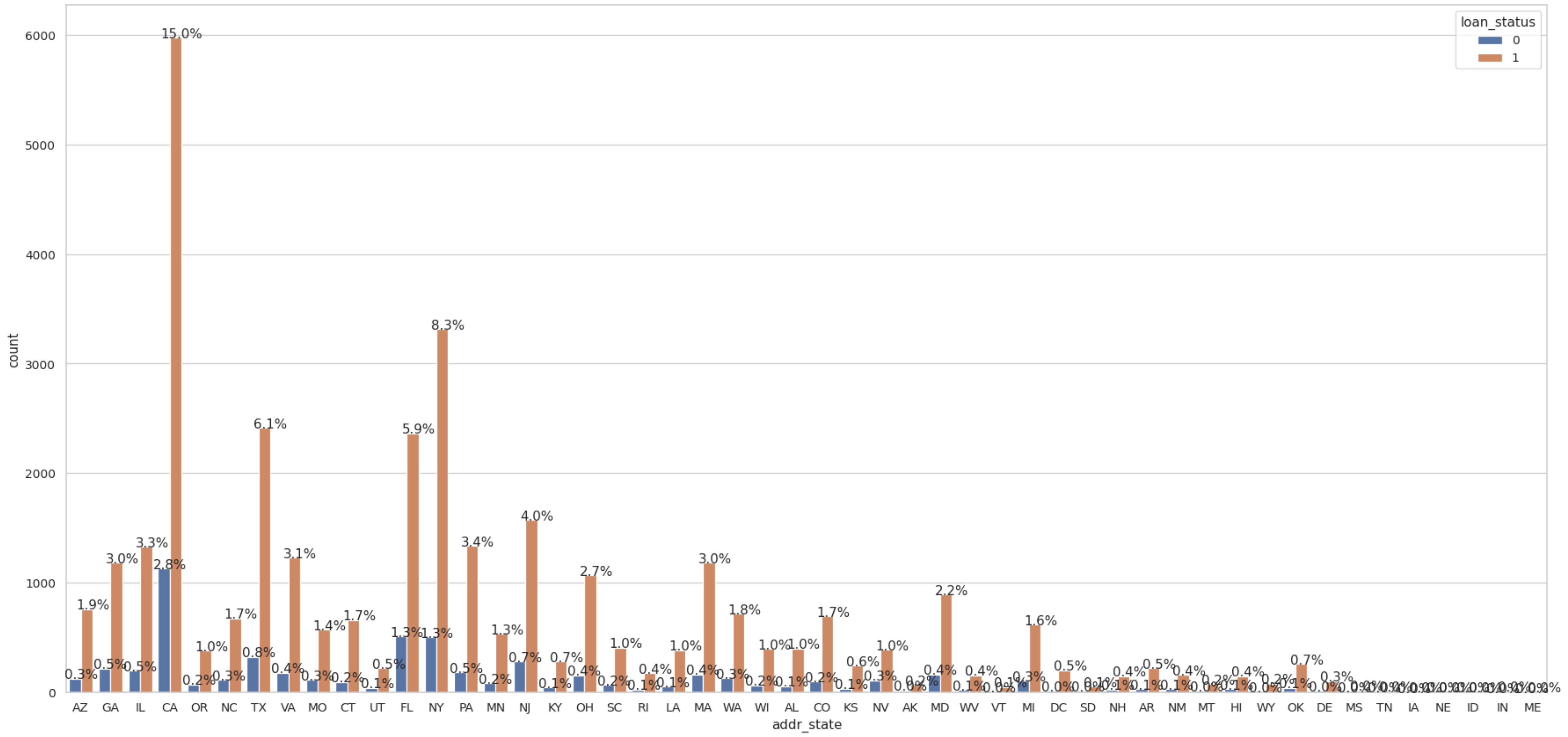
'F' and 'G' sub-grades don't get paid back that often



Better to avoid small business loans followed by educational loans



The CA, NY, TX states had high number of applications and high default rate



Machine Learning based Loan Default Predictions

Modeling Binary Classifiers

Five binary classifiers have been modeled namely, Linear SVC, Logistic Regression, Gaussian NB, Random Forest Classier, Gradient Boosting Classifier and XGBClassifier

Chosen Recall, Precision, and F1-score as evaluation metrics.

The precision is the measure of how accurate the classifier's prediction of a specific class.

The Recall is the measure of the classifier's ability to identify a class.

Resampling (Oversampling)

> This technique is used to upsample the minority class of an imbalanced dataset using replacement. This technique is called oversampling.

Synthetic Minority Oversampling Technique (SMOTE)

> SMOTE is another technique to oversample the minority class. It looks into minority class instances and uses k nearest neighbor to pick a random nearest neighbor, and a synthetic instance is created randomly in feature space.

Classification Metrics without application of Data Imbalance handling Techniques

Linear SVC

	precision	recall	f1-score	support
0.0	0.33	0.00	0.00	534
1.0	0.87	1.00	0.93	3440
accuracy			0.87	3974
macro avg	0.60	0.50	0.47	3974
weighted avg	0.79	0.87	0.80	3974

Logistic Regression

	precision	recall	f1-score	support
0.0	0.25	0.00	0.01	534
1.0	0.87	1.00	0.93	3440
accuracy			0.86	3974
macro avg	0.56	0.50	0.47	3974
weighted avg	0.78	0.86	0.80	3974

Gaussian NB

	precision	recall	f1-score	support
0.0	0.25	0.32	**0.28**	534
1.0	0.89	0.85	**0.87**	3440
accuracy			0.78	3974
macro avg	0.57	0.59	0.58	3974
weighted avg	0.80	0.78	0.79	3974

RandomForest Classifier

	precision	recall	f1-score	support
0.0	0.33	0.10	0.16	534
1.0	0.87	0.97	0.92	3440
accuracy			0.85	3974
macro avg	0.60	0.54	0.54	3974
weighted avg	0.80	0.85	0.82	3974

GradientBoosting Classifier

	precision	recall	f1-score	support
0.0	0.00	0.00	0.00	534
1.0	0.87	1.00	0.93	3440
accuracy			0.87	3974
macro avg	0.43	0.50	0.46	3974
weighted avg	0.75	0.87	0.80	3974

XGBClassifier

	precision	recall	f1-score	support
0.0	0.38	0.06	0.11	534
1.0	0.87	0.98	0.92	3440
accuracy			0.86	3974
macro avg	0.62	0.52	0.52	3974
weighted avg	0.80	0.86	0.81	3974

Classification Metrics with application of Data Imbalance handling Techniques-Resampling (Oversampling)

Linear SVC

	precision	recall	f1-score	support
0.0	0.56	0.68	0.62	3367
1.0	0.61	0.48	0.54	3450
accuracy			0.58	6817
macro avg	0.59	0.58	0.58	6817
weighted avg	0.59	0.58	0.58	6817

RandomForest Classifier

	precision	recall	f1-score	support
0.0	0.80	0.94	0.86	3367
1.0	0.93	0.76	0.84	3450
accuracy			0.85	6817
macro avg	0.86	0.85	0.85	6817
weighted avg	0.87	0.85	0.85	6817

Logistic Regression

	precision	recall	f1-score	support
0.0	0.63	0.57	0.60	3367
1.0	0.61	0.67	0.64	3450
accuracy			0.62	6817
macro avg	0.62	0.62	0.62	6817
weighted avg	0.62	0.62	0.62	6817

GradientBoosting Classifier

	precision	recall	f1-score	support
0.0	0.62	0.66	0.64	3367
1.0	0.65	0.61	0.63	3450
accuracy			0.63	6817
macro avg	0.63	0.63	0.63	6817
weighted avg	0.63	0.63	0.63	6817

Gaussian NB

	precision	recall	f1-score	support
0.0	0.62	0.59	0.60	3367
1.0	0.61	0.64	0.63	3450
accuracy			0.61	6817
macro avg	0.61	0.61	0.61	6817
weighted avg	0.61	0.61	0.61	6817

XGBClassifier

	precision	recall	f1-score	support
0.0	0.74	0.82	0.78	3367
1.0	0.80	0.72	0.76	3450
accuracy			0.77	6817
macro avg	0.77	0.77	0.77	6817
weighted avg	0.77	0.77	0.77	6817

Classification Metrics with application of Data Imbalance handling Techniques-SMOTE

Linear SVC

	precision	recall	f1-score	support
0.0	0.51	1.00	0.68	3479
1.0	0.22	0.00	0.00	3338
accuracy			0.51	6817
macro avg	0.37	0.50	0.34	6817
weighted avg	0.37	0.51	0.35	6817

Logistic Regression

	precision	recall	f1-score	support
0.0	0.64	0.57	0.60	3479
1.0	0.59	0.66	0.63	3338
accuracy			0.61	6817
macro avg	0.62	0.61	0.61	6817
weighted avg	0.62	0.61	0.61	6817

Gaussian NB

	precision	recall	f1-score	support
0.0	0.61	0.74	0.67	3479
1.0	0.65	0.50	0.57	3338
accuracy			0.62	6817
macro avg	0.63	0.62	0.62	6817
weighted avg	0.63	0.62	0.62	6817

RandomForest Classifier

	precision	recall	f1-score	support
0.0	0.94	0.85	**0.89**	3479
1.0	0.85	0.94	**0.90**	3338
accuracy			0.89	6817
macro avg	0.90	0.89	0.89	6817
weighted avg	0.90	0.89	0.89	6817

GradientBoosting Classifier

	precision	recall	f1-score	support
0.0	0.85	0.79	0.82	3479
1.0	0.80	0.86	0.83	3338
accuracy			0.82	6817
macro avg	0.83	0.82	0.82	6817
weighted avg	0.83	0.82	0.82	6817

XGBClassifier

	precision	recall	f1-score	support
0.0	0.98	0.84	**0.90**	3479
1.0	0.85	0.98	**0.91**	3338
accuracy			0.91	6817
macro avg	0.92	0.91	0.91	6817
weighted avg	0.92	0.91	0.91	6817

Machine Learning based Loan Default Predictions

Modeling Binary Classifiers after Data Augmentation

New Dataset-Classification Metrics without application of Data Imbalance handling Techniques

Linear SVC

```
              precision    recall  f1-score   support
           0       0.82      0.86      0.84      9674
           1       0.96      0.94      0.95     32378
    accuracy                           0.93     42052
   macro avg       0.89      0.90      0.90     42052
weighted avg       0.93      0.93      0.93     42052
```

Logistic Regression

```
              precision    recall  f1-score   support
           0       0.87      0.83      0.85      9674
           1       0.95      0.96      0.96     32378
    accuracy                           0.93     42052
   macro avg       0.91      0.90      0.91     42052
weighted avg       0.93      0.93      0.93     42052
```

Gaussian NB

```
              precision    recall  f1-score   support
           0       0.58      0.74      0.65      9674
           1       0.91      0.84      0.87     32378
    accuracy                           0.81     42052
   macro avg       0.74      0.79      0.76     42052
weighted avg       0.84      0.81      0.82     42052
```

RandomForest Classifier

```
              precision    recall  f1-score   support
           0       0.84      0.87      0.86      9674
           1       0.96      0.95      0.96     32378
    accuracy                           0.93     42052
   macro avg       0.90      0.91      0.91     42052
weighted avg       0.93      0.93      0.93     42052
```

GradientBoosting Classifier

```
              precision    recall  f1-score   support
           0       0.84      0.88      0.86      9674
           1       0.96      0.95      0.96     32378
    accuracy                           0.94     42052
   macro avg       0.90      0.92      0.91     42052
weighted avg       0.94      0.94      0.94     42052
```

XGBClassifier

```
              precision    recall  f1-score   support
           0       0.86      0.87      0.86      9674
           1       0.96      0.96      0.96     32378
    accuracy                           0.94     42052
   macro avg       0.91      0.92      0.91     42052
weighted avg       0.94      0.94      0.94     42052
```

New Dataset-Classification Metrics with application of Data Imbalance handling Techniques-Resampling (Oversampling)

Linear SVC

	precision	recall	f1-score	support
0	0.90	0.89	0.90	32315
1	0.89	0.90	0.90	32462
accuracy			0.90	64777
macro avg	0.90	0.90	0.90	64777
weighted avg	0.90	0.90	0.90	64777

Logistic Regression

	precision	recall	f1-score	support
0	0.93	0.92	**0.92**	32315
1	0.92	0.93	**0.93**	32462
accuracy			0.92	64777
macro avg	0.92	0.92	0.92	64777
weighted avg	0.92	0.92	0.92	64777

Gaussian NB

	precision	recall	f1-score	support
0	0.79	0.92	0.85	32315
1	0.91	0.75	0.82	32462
accuracy			0.84	64777
macro avg	0.85	0.84	0.84	64777
weighted avg	0.85	0.84	0.84	64777

RandomForest Classifier

	precision	recall	f1-score	support
0	0.93	0.98	**0.95**	32315
1	0.98	0.93	**0.95**	32462
accuracy			0.95	64777
macro avg	0.95	0.95	0.95	64777
weighted avg	0.95	0.95	0.95	64777

GradientBoosting Classifier

	precision	recall	f1-score	support
0	0.92	0.93	0.92	32315
1	0.93	0.92	0.92	32462
accuracy			0.92	64777
macro avg	0.92	0.92	0.92	64777
weighted avg	0.92	0.92	0.92	64777

XGBClassifier

	precision	recall	f1-score	support
0	0.93	0.94	**0.93**	32315
1	0.94	0.93	**0.93**	32462
accuracy			0.93	64777
macro avg	0.93	0.93	0.93	64777
weighted avg	0.93	0.93	0.93	64777

New Dataset-Classification Metrics with application of Data Imbalance handling Techniques-SMOTE

Linear SVC

```
              precision    recall  f1-score   support
           0       0.88      0.96      0.92     32432
           1       0.96      0.86      0.91     32345
    accuracy                           0.91     64777
   macro avg       0.92      0.91      0.91     64777
weighted avg       0.92      0.91      0.91     64777
```

RandomForest Classifier

```
              precision    recall  f1-score   support
           0       0.94      0.97      0.95     32432
           1       0.97      0.94      0.95     32345
    accuracy                           0.95     64777
   macro avg       0.95      0.95      0.95     64777
weighted avg       0.95      0.95      0.95     64777
```

Logistic Regression

```
              precision    recall  f1-score   support
           0       0.94      0.94      0.94     32432
           1       0.94      0.94      0.94     32345
    accuracy                           0.94     64777
   macro avg       0.94      0.94      0.94     64777
weighted avg       0.94      0.94      0.94     64777
```

GradientBoosting Classifier

```
              precision    recall  f1-score   support
           0       0.94      0.95      0.94     32432
           1       0.95      0.93      0.94     32345
    accuracy                           0.94     64777
   macro avg       0.94      0.94      0.94     64777
weighted avg       0.94      0.94      0.94     64777
```

Gaussian NB

```
              precision    recall  f1-score   support
           0       0.87      0.91      0.89     32432
           1       0.91      0.86      0.88     32345
    accuracy                           0.89     64777
   macro avg       0.89      0.89      0.89     64777
weighted avg       0.89      0.89      0.89     64777
```

XGBClassifier

```
              precision    recall  f1-score   support
           0       0.96      0.96      0.96     32432
           1       0.96      0.95      0.96     32345
    accuracy                           0.96     64777
   macro avg       0.96      0.96      0.96     64777
weighted avg       0.96      0.96      0.96     64777
```

ANN Classification Metrics ATT LC Dataset Vs. Augmented Dataset

	precision	recall	f1-score	support
0.0	0.25	0.00	0.01	534
1.0	0.87	1.00	0.93	3440
accuracy			0.86	3974
macro avg	0.56	0.50	0.47	3974
weighted avg	0.78	0.86	0.80	3974

	precision	recall	f1-score	support
0	0.86	0.85	**0.85**	9674
1	0.96	0.96	**0.96**	32378
accuracy			0.93	42052
macro avg	0.91	0.90	0.91	42052
weighted avg	0.93	0.93	0.93	42052

Conclusions

Driving Factors (or driver variables)

(1) Grade: Default Rate is high in high risk loan applicants. It is important to thoroughly check high risk loan applications.

(2) Installment Amount: Defaulter rate increases as the requested loan installment amount increases.

(4) Annual Income: Applicants from Low income group have a greater share of defaulted loans.

(5) **States:** The CA, NY, TX states had high number of applications and high default rate. The plot
is represented in Figure 16.

(6) **Purpose:** Better to avoid small business loans followed by educational loans

(7) **DTI**: Higher DTI translates to higher default rates

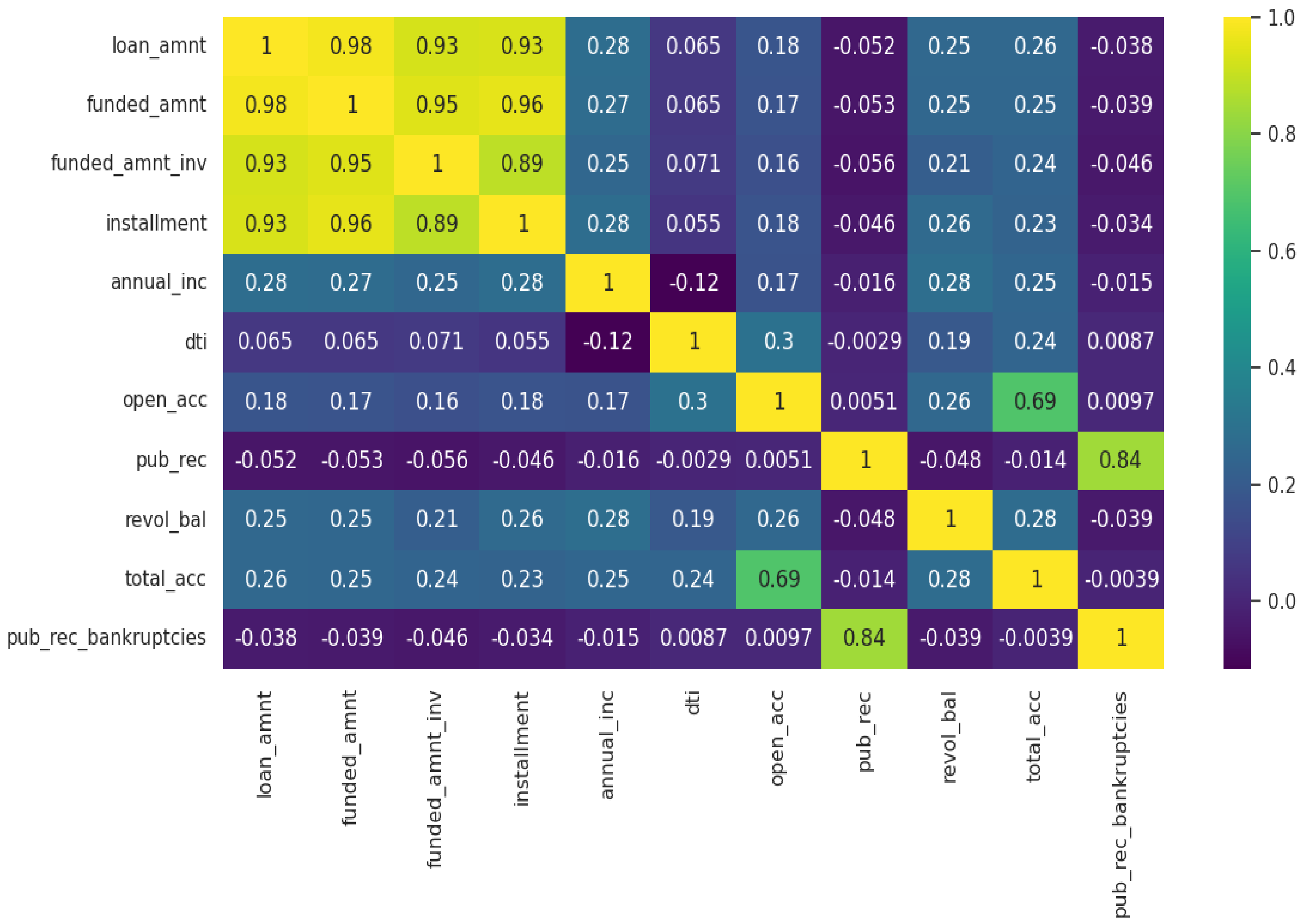
(8) **Income source verification:** Should be checked thoroughly.

(9) **Binary Classification:** Random Forest and XGBoost provided the best F-Scores and produces good accuracy as well.

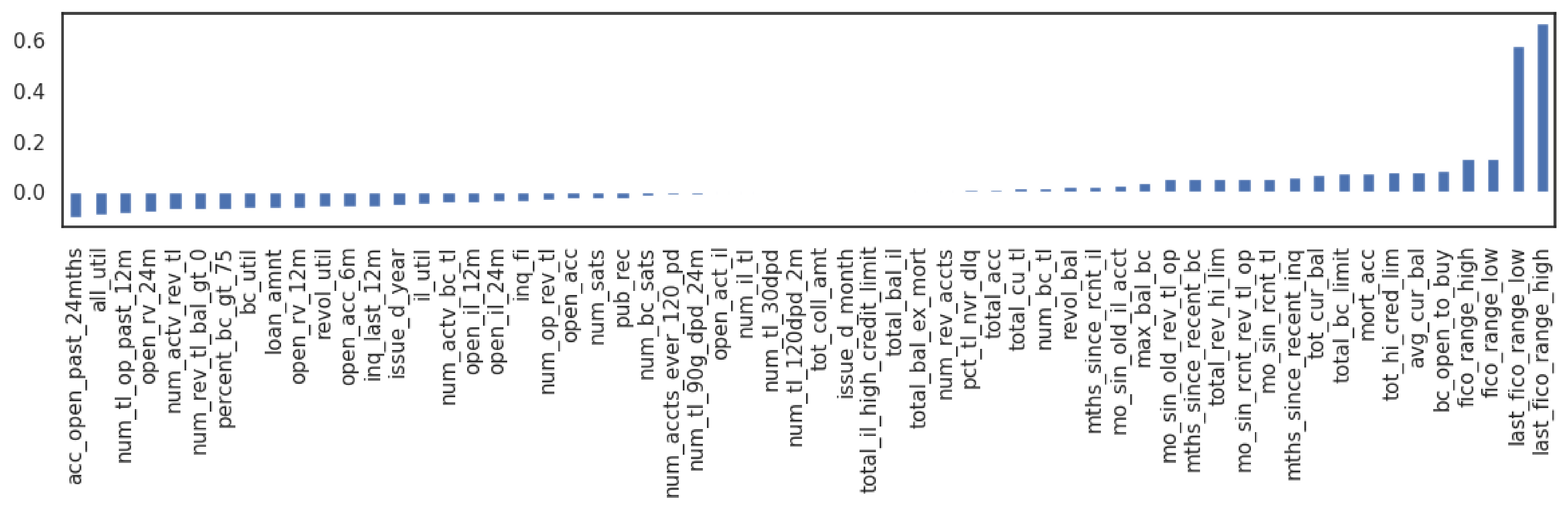
(10) **ANN:** ANN F-score improved on Augmented data.

Thank you

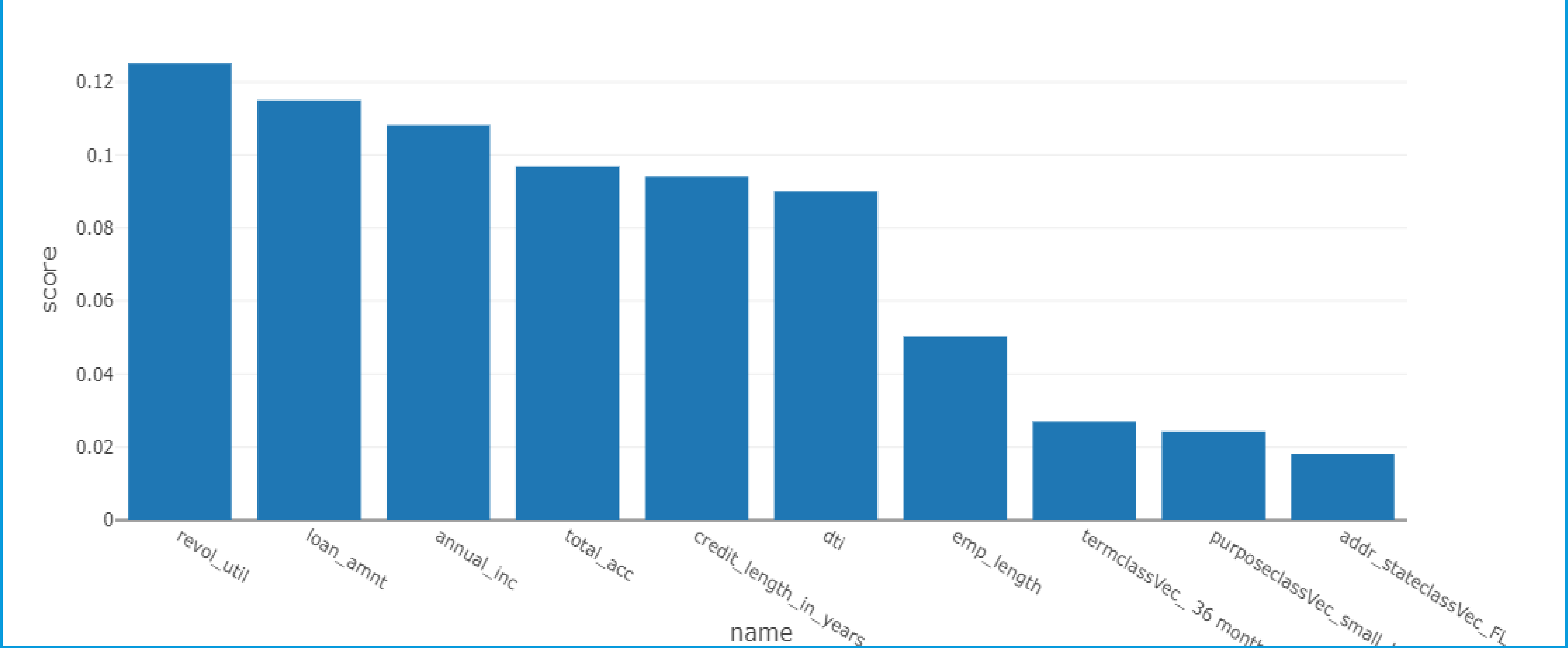
LC Loan Stats



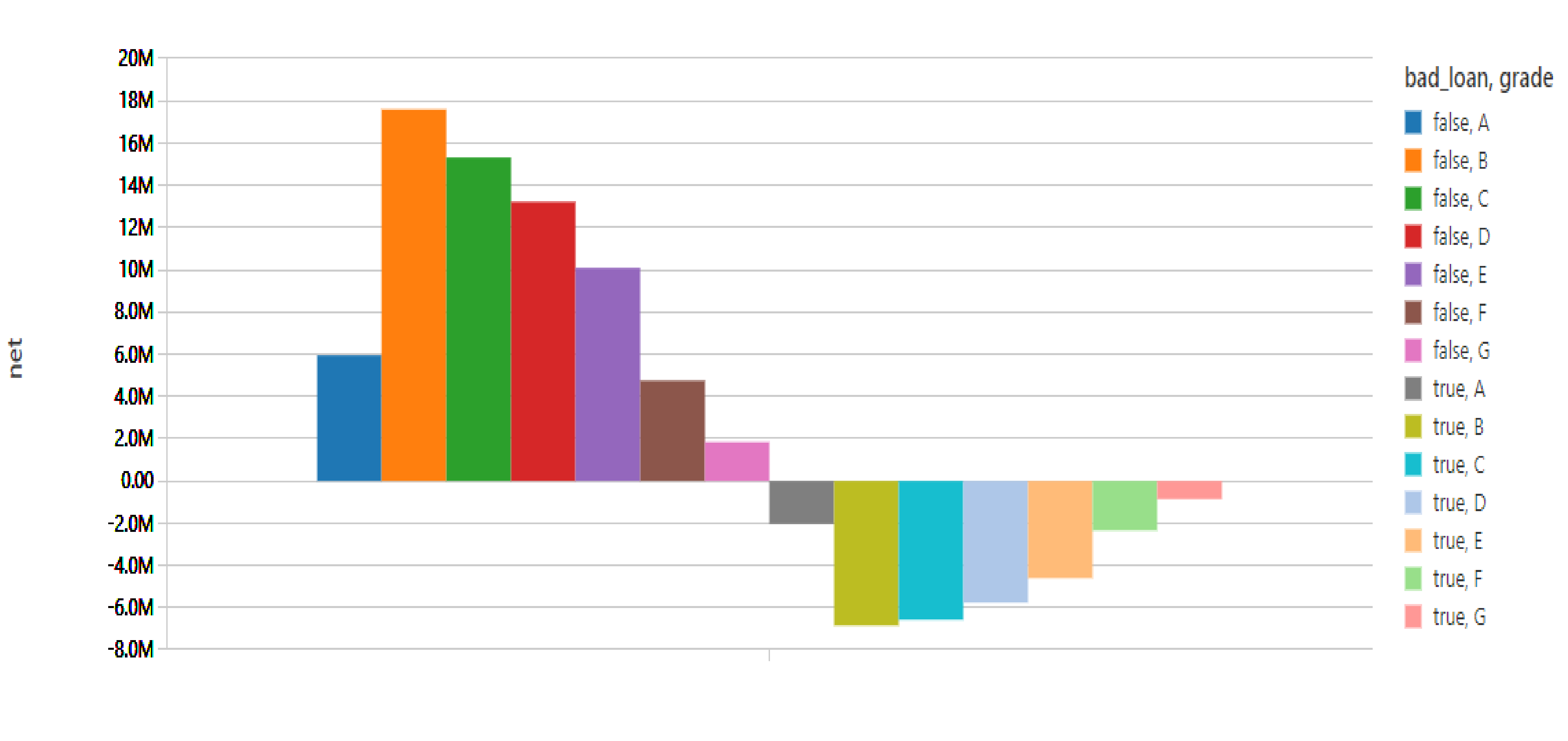
	loan_amnt	funded_amnt	funded_amnt_inv	installment	annual_inc	dti	open_acc	pub_rec	revol_bal	total_acc	pub_rec_bankruptcies
loan_amnt	1	0.98	0.93	0.93	0.28	0.065	0.18	-0.052	0.25	0.26	-0.038
funded_amnt	0.98	1	0.95	0.96	0.27	0.065	0.17	-0.053	0.25	0.25	-0.039
funded_amnt_inv	0.93	0.95	1	0.89	0.25	0.071	0.16	-0.056	0.21	0.24	-0.046
installment	0.93	0.96	0.89	1	0.28	0.055	0.18	-0.046	0.26	0.23	-0.034
annual_inc	0.28	0.27	0.25	0.28	1	-0.12	0.17	-0.016	0.28	0.25	-0.015
dti	0.065	0.065	0.071	0.055	-0.12	1	0.3	-0.0029	0.19	0.24	0.0087
open_acc	0.18	0.17	0.16	0.18	0.17	0.3	1	0.0051	0.26	0.69	0.0097
pub_rec	-0.052	-0.053	-0.056	-0.046	-0.016	-0.0029	0.0051	1	-0.048	-0.014	0.84
revol_bal	0.25	0.25	0.21	0.26	0.28	0.19	0.26	-0.048	1	0.28	-0.039
total_acc	0.26	0.25	0.24	0.23	0.25	0.24	0.69	-0.014	0.28	1	-0.0039
pub_rec_bankruptcies	-0.038	-0.039	-0.046	-0.034	-0.015	0.0087	0.0097	0.84	-0.039	-0.0039	1


0.6
0.4
0.2
0.0
acc_open_past_24mths
all_util
num_tl_op_past_12m
open_rv_24m
num_actv_rev_tl
num_rev_tl_bal_gt_0
percent_bc_gt_75
bc_util
loan_amnt
open_rv_12m
revol_util
open_acc_6m
inq_last_12m
issue_d_year
il_util
num_actv_bc_tl
open_il_12m
open_il_24m
inq_fi
num_op_rev_tl
open_acc
num_sats
pub_rec
num_bc_sats
num_accts_ever_120_pd
num_tl_90g_dpd_24m
open_act_il
num_il_tl
num_tl_30dpd
num_tl_120dpd_2m
tot_coll_amt
issue_d_month
total_il_high_credit_limit
total_bal_il
total_bal_ex_mort
num_rev_accts
pct_tl_nvr_dlq
total_acc
total_cu_tl
num_bc_tl
revol_bal
mths_since_rcnt_il
mo_sin_old_il_acct
max_bal_bc
mo_sin_old_rev_tl_op
mths_since_recent_bc
total_rev_hi_lim
mo_sin_rcnt_rev_tl_op
mo_sin_rcnt_tl
mths_since_recent_inq
tot_cur_bal
total_bc_limit
mort_acc
tot_hi_cred_lim
avg_cur_bal
bc_open_to_buy
fico_range_high
fico_range_low
last_fico_range_low
last_fico_range_high

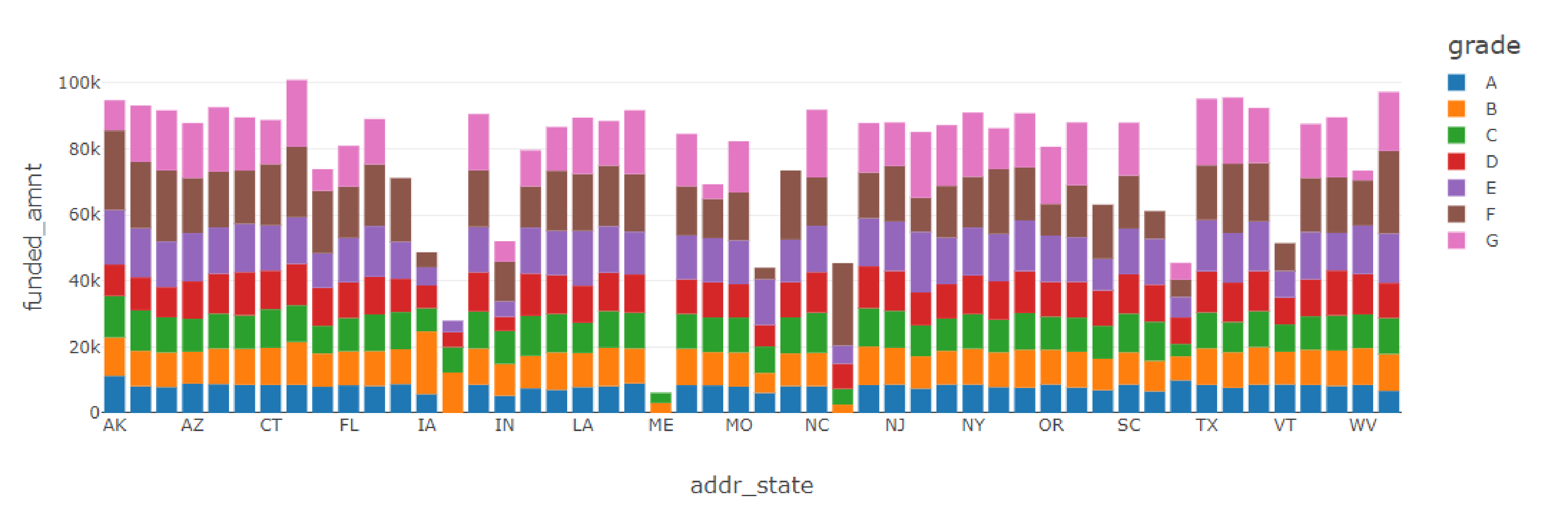
Identify Important Features from Model

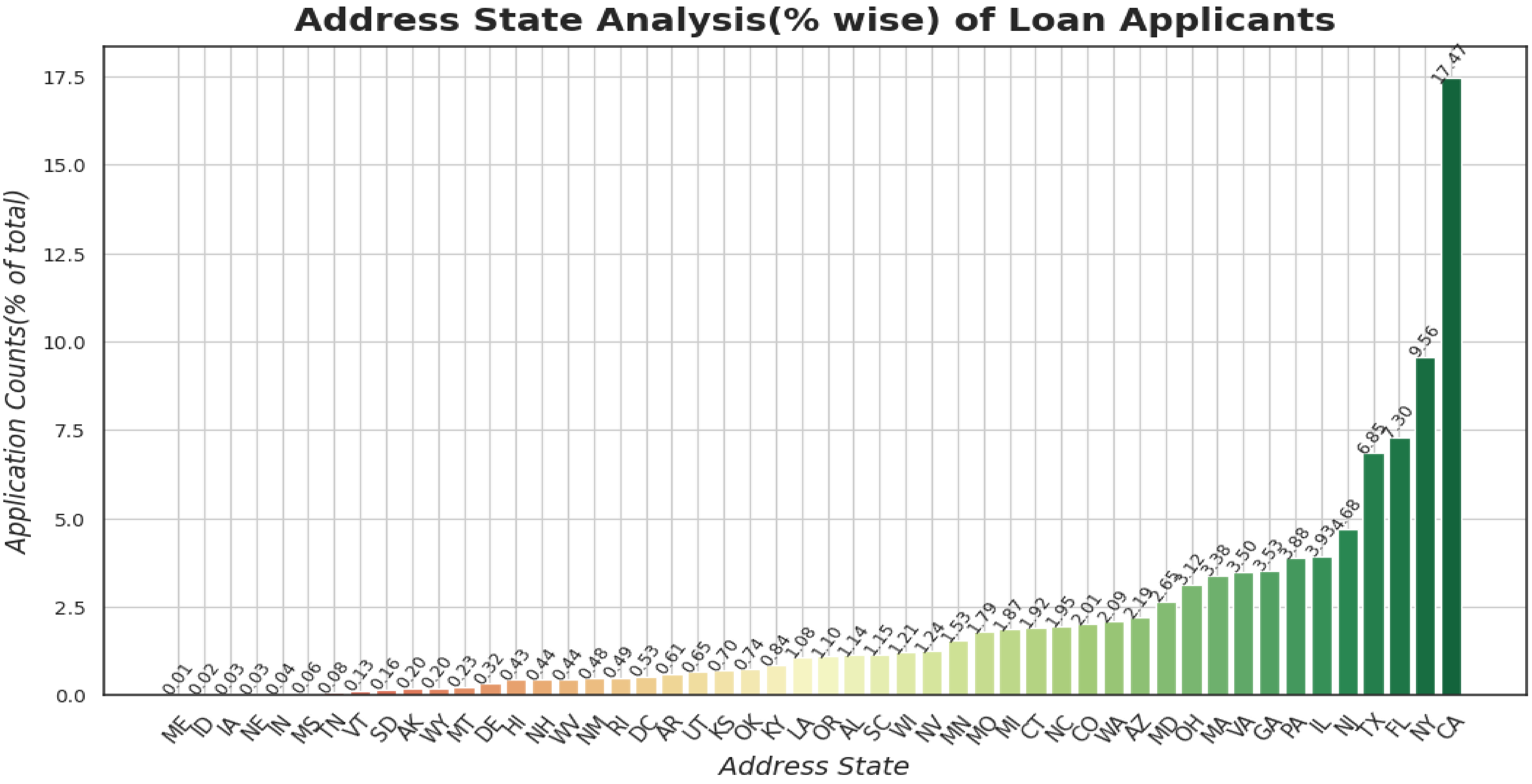


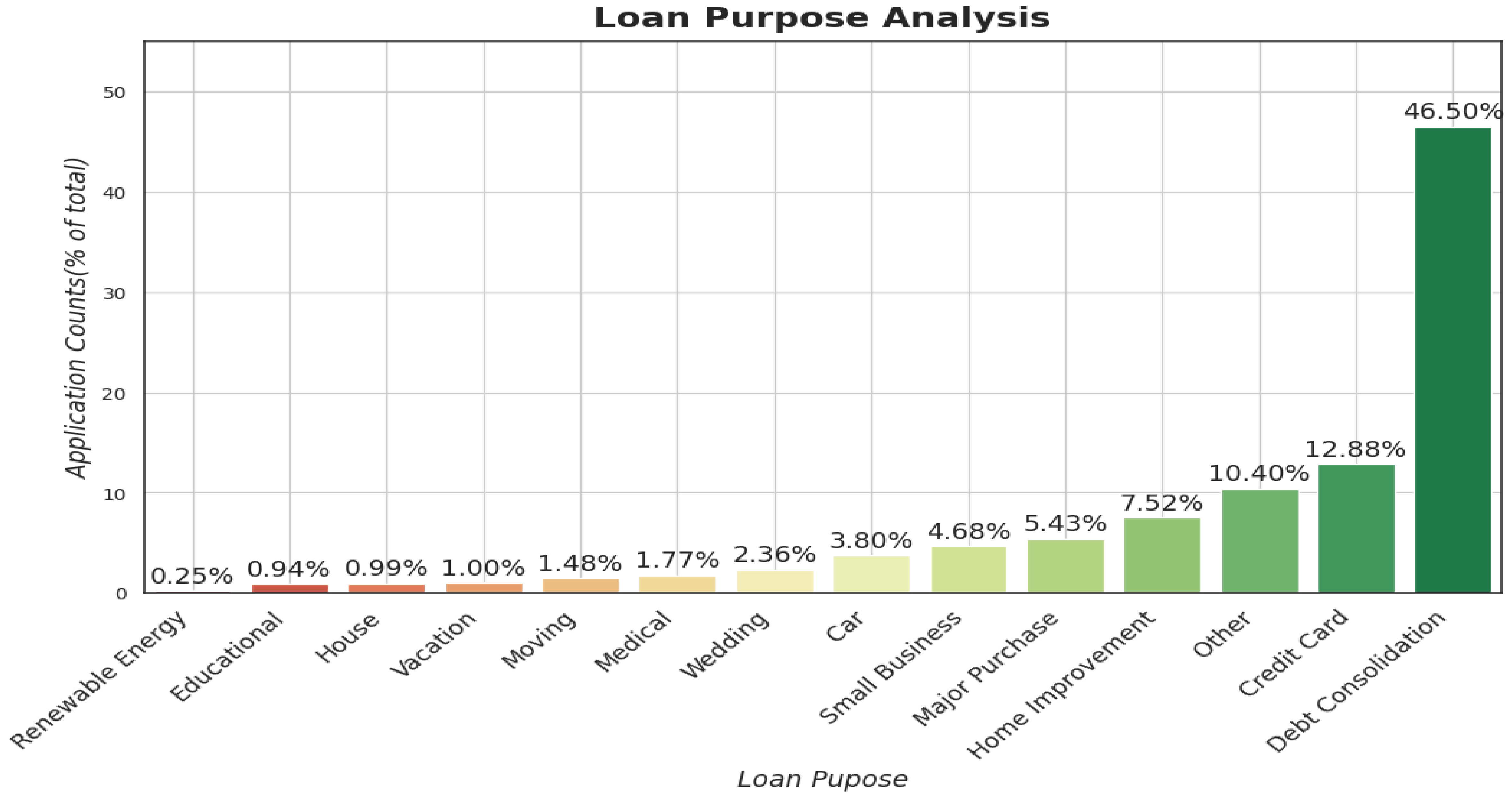
Net Profits by Asset Class and Default Status



LC Asset Allocation by Grade




Address State Analysis(% wise) of Loan Applicants
Application Counts(% of total)
Address State
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
ME 0.01
ID 0.02
IA 0.03
NE 0.03
IN 0.04
MS 0.06
TN 0.08
VT 0.13
SD 0.16
AK 0.20
WY 0.20
MT 0.23
DE 0.32
HI 0.43
NH 0.44
WV 0.44
NM 0.48
RI 0.49
DC 0.53
AR 0.61
UT 0.65
KS 0.70
OK 0.74
KY 0.84
LA 1.08
OR 1.10
AL 1.14
SC 1.15
WI 1.21
NV 1.24
MN 1.53
MO 1.79
MI 1.87
CT 1.92
NC 1.95
CO 2.01
WA 2.09
AZ 2.19
MD 2.65
OH 3.12
MA 3.38
VA 3.50
GA 3.53
PA 3.88
IL 3.93
NJ 4.68
TX 6.85
FL 7.30
NY 9.56
CA 17.47


Loan Purpose Analysis
Application Counts(% of total)
0
10
20
30
40
50
0.25%
0.94%
0.99%
1.00%
1.48%
1.77%
2.36%
3.80%
4.68%
5.43%
7.52%
10.40%
12.88%
46.50%
Renewable Energy
Educational
House
Vacation
Moving
Medical
Wedding
Car
Small Business
Major Purchase
Home Improvement
Other
Credit Card
Debt Consolidation
Loan Pupose