LendingClub Loan Default Assessment Notes

Rama Reddy R

April 6, 2023

1

Find the average annual income and number of loan applicants by each state .

Solution

Average annual income and number of applicants by State

The corresponding graphical representations are given in Figure1 and Figure 2 respectively

State	Applicants	Average Income
AK	86.0	78902.430698
AL	484.0	63275.835661
AR	261.0	59946.404674
AZ	933.0	67799.876420
CA	7429.0	72221.438542
CO	857.0	66823.789627
CT	816.0	75707.016336
DC	224.0	77794.437500
DE	136.0	69437.426471
FL	3104.0	64789.772764
GA	1503.0	69255.655868
HI	181.0	64575.228950
IA	12.0	45548.916667
ID	9.0	47569.733333
IL	1672.0	69790.300377
IN	19.0	32733.526316
KS	298.0	62332.194899
KY	359.0	61349.168217
LA	461.0	73649.697354
MA	1438.0	72877.093700
MD	1125.0	78283.814551
ME	3.0	23866.666667

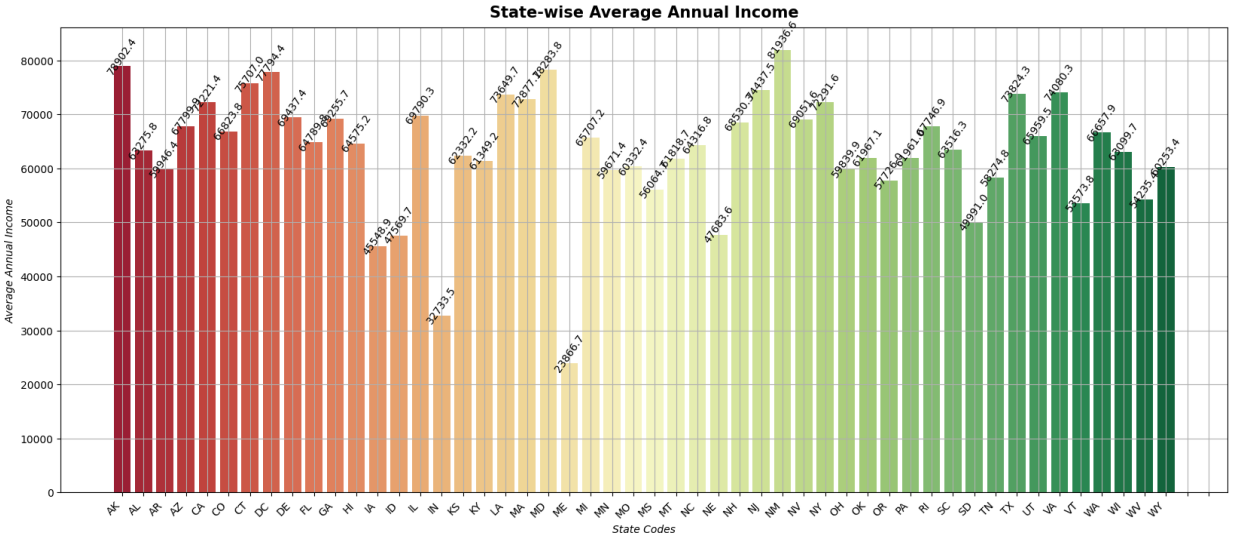

Figure 1: State-wise average annual income

MI	796.0	65707.246457
MN	652.0	59671.433206
MO	765.0	60332.378379
MS	26.0	56064.653846
MT	96.0	61818.727083
NC	830.0	64316.825627
NE	11.0	47683.636364
NH	188.0	68530.272447
NJ	1988.0	74437.502384
NM	205.0	81936.604488
NV	527.0	69051.606869
NY	4061.0	72291.583960
OH	1329.0	59839.882611
OK	317.0	61967.074322
OR	468.0	57726.024915
PA	1651.0	61960.990860
RI	208.0	67746.926923
SC	489.0	63516.255542
SD	67.0	49990.969552
TN	32.0	58274.812500
TX	2915.0	73824.332978
UT	278.0	65959.534964
VA	1487.0	74080.299657
VT	57.0	53573.782632
WA	888.0	66657.879854
WI	516.0	63099.688198
WV	187.0	54235.402941
WY	87.0	60253.426207

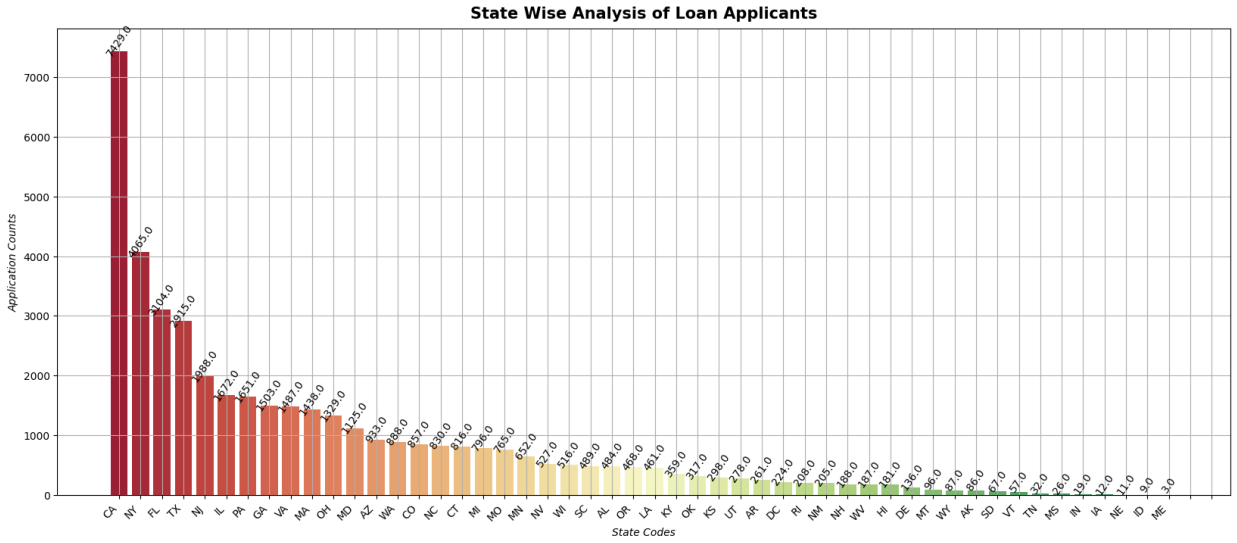

Figure 2: Number of Loan applicants from each state

:

2

We're interested in **predicting which applicants will default** on their loan.

a. **Build a derived variable** representing whether an applicant defaulted on their loan. Consider a loan that is "Charged Off" as a default, and a loan that is "Fully Paid" to not have defaulted.

b. Build a **binary classification model** to predict which loans will default. Your model should take a data set of applicants as input and return the probability of default for each applicant. You should thoroughly describe how you developed and validated your model and explain any assumptions you made.

c. Imagine your client is considering entering the lending market but is very risk averse (they prefer low default rates even if it means accepting lower rates of return). **Develop a strategy** for entering this market.

Some things to consider:

i. Which locations should we target?

ii. To which segments of the population should we advertise?

iii. Any other helpful strategies you can think of to keep default rates low?

Solution:-

a. **Build a derived variable** representing whether an applicant defaulted on their loan. Consider a loan that is "Charged Off" as a default, and a loan that is "Fully Paid" to not have defaulted.

Exploratory Data Analysis (EDA) was performed to obtain the derived variables for assessing the defaulted loans.

Data wrangling was performed to identify the predictors of default for approval/rejection of the loan so that the risk is minimized.

The Lending Club data majorly consists of 3 types of variables -

Applicant demographic variables such occupation, employment details etc. Loan metrics such as amount of loan, interest rate, purpose of loan etc. Customer actions variables such as delinquent 2 years, revolving balance, next payment date etc.

As the customer actions variables will not be available at the time of loan application, and hence those can not be used as possible predictors for loan default.

Going forward, analysis uses only the other two types of variables.

Further, the correlation matrix of numeric catergories has been analyzed to check for possible load default predictors. The correlation matrix is given in Figure 3.

And, the distribution of target variable is depicted in Figure 4. And this represents a perfect "class imbalance" classification problem.

Observations from correlation matrix, statistics on loan amount, and other categories versus loan status has been highlighted in the following sections

There is a class imbalance. we have more cases of 'Fully Paid (encoded as '1') than the 'Charged Off'(encoded as '0')

There is not a much difference between the defaulters w.r.t to loan amount perse.

The loan installment is correlated to to loan amount and the reason being loan installment is calculated based on loan amount.

The 'E' and 'F' grade loand are defaulted more often and also 'F' and 'G' sub-grades don't get paid back that often.The plots are given in Figure 5 and Figure 6.

The Annual income is one of the important factor and in deciding factor in loan repayment as shown in the correlation plot in Figure 7.

Figure 3: Correlation matrix of numerical categories

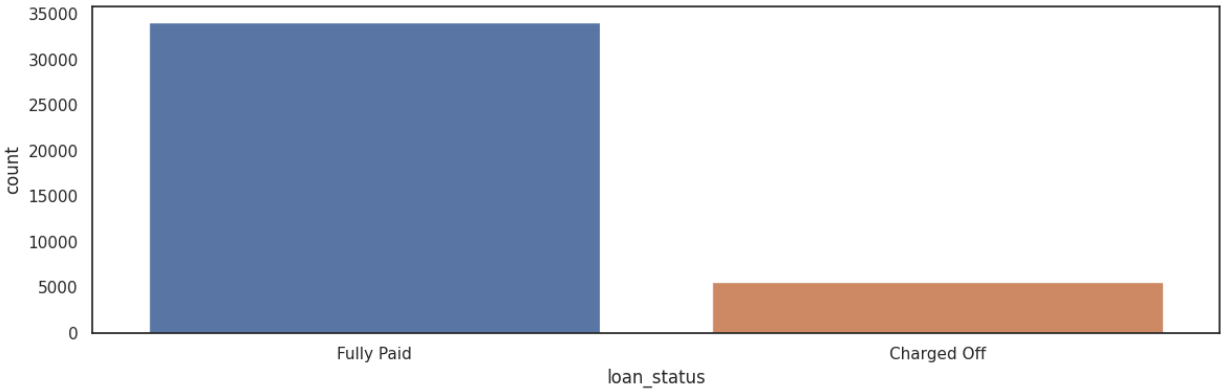

Figure 4: Distribution of target variable

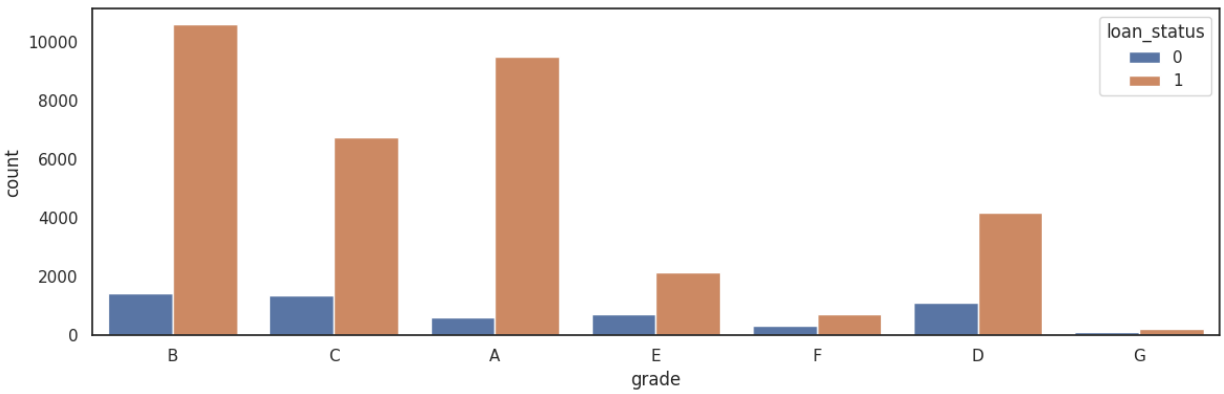

Figure 5: Grade vs.Loan status

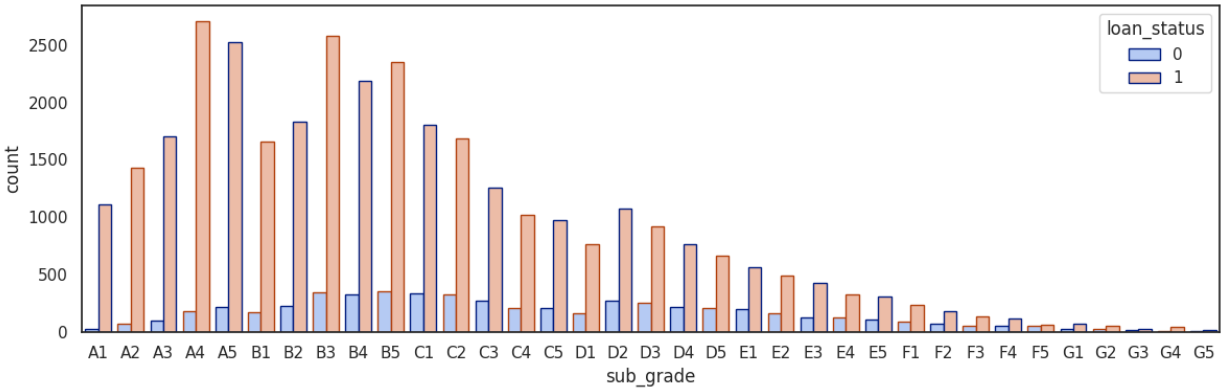

Figure 6: sub grade vs. Loan status

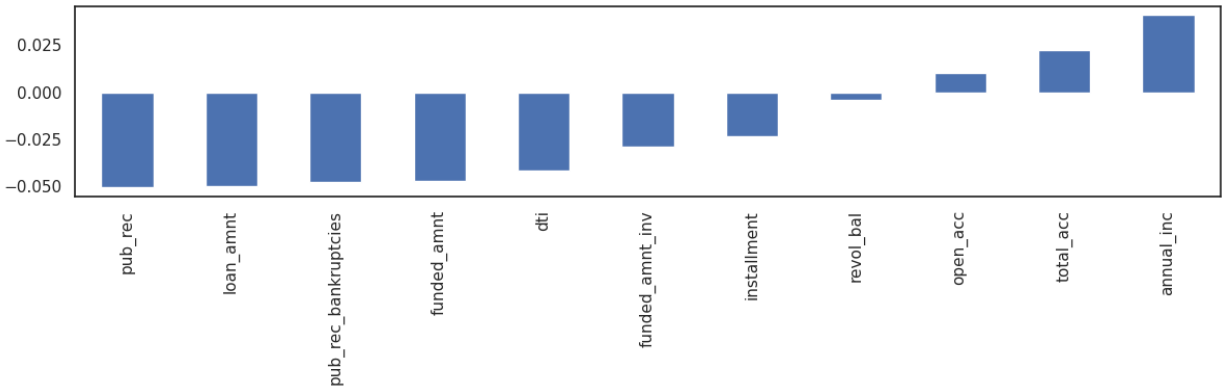

Figure 7: Annual income vs loan status

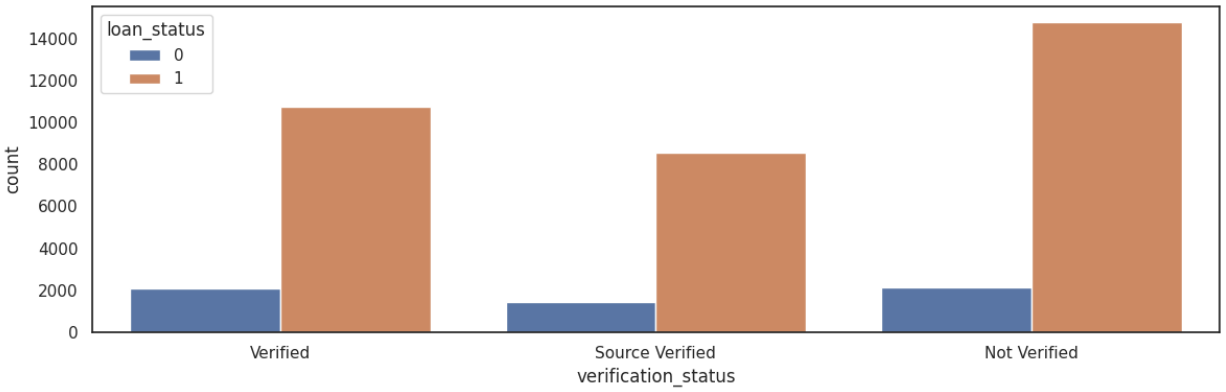

Figure 8: Verification status vs loan status

The more verified people defaulted than the not verified which calls for better verification process as represented in Figure 8.

The small business loans default the most, then the renewable energy and education loans as shown in Figure 9.

The 60 months loans default more than 36 months loans as shown in Figure 10.

High interest rates loans are defaulted the most as represented in Figure 11.

Comparing default rates across debt to income ratio(dti), high dti translates into higher default rates, as expected, shown in Figure 12.

Comparing default rates across installment, higher the installment amount, higher the default rate as shown in Figure 13

b. Build a binary classification model to predict which loans will default. Your model should take a data set of applicants as input and return the prob-

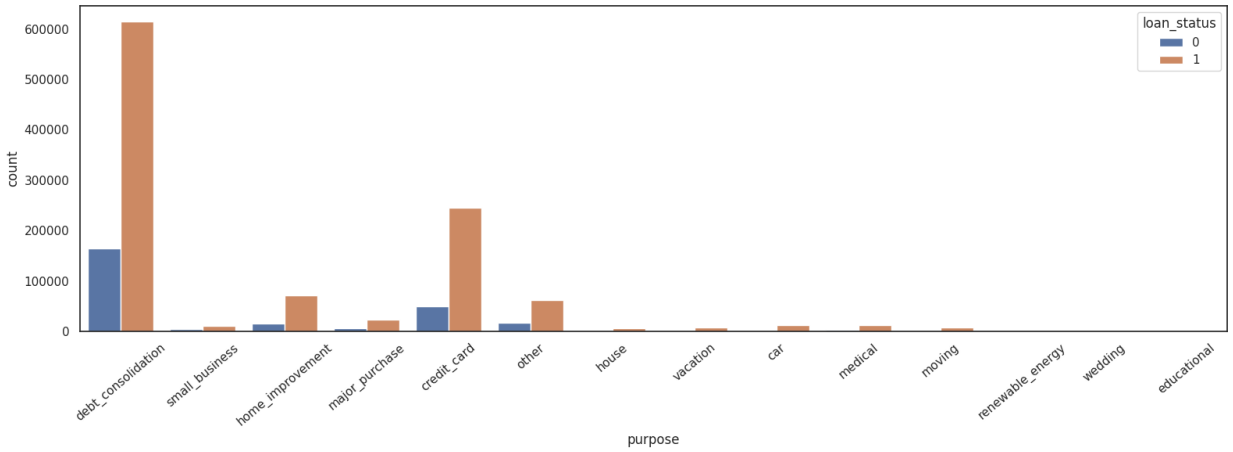

Figure 9: purpose vs. loan status

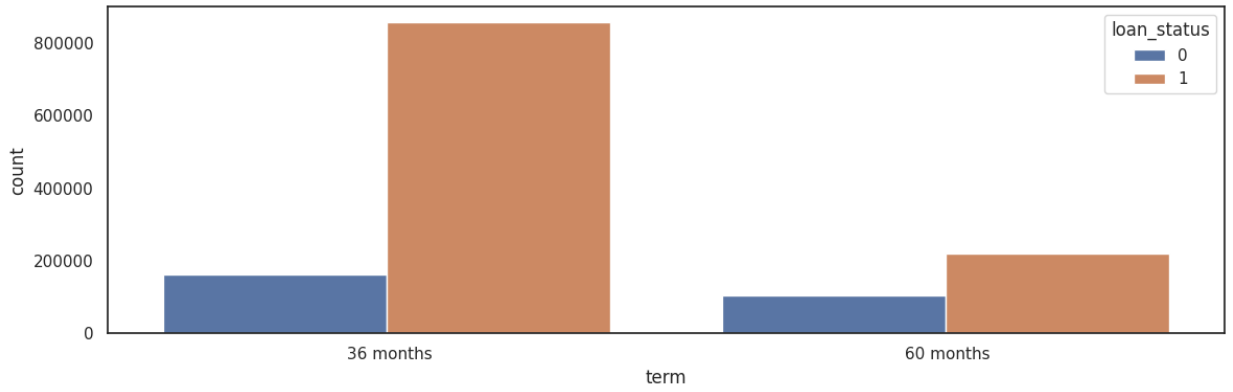

Figure 10: term vs. loan status

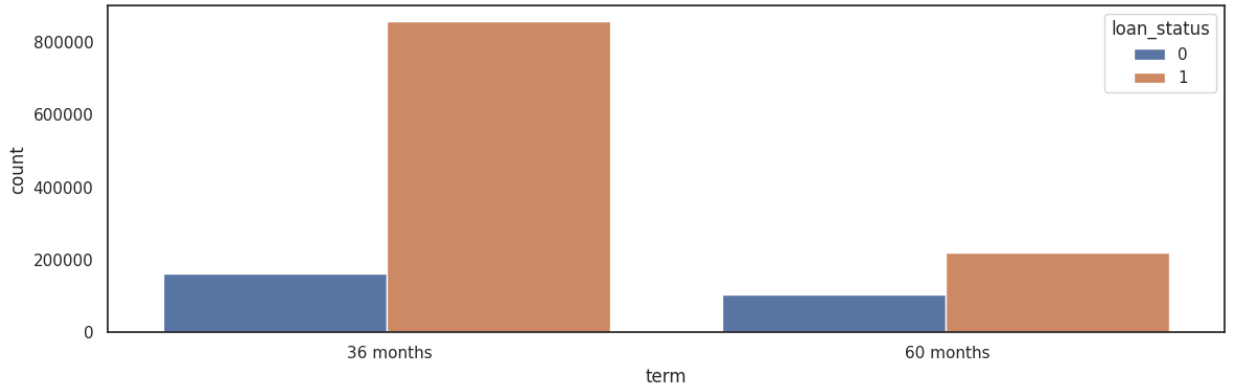

Figure 11: interest rate vs. loan status

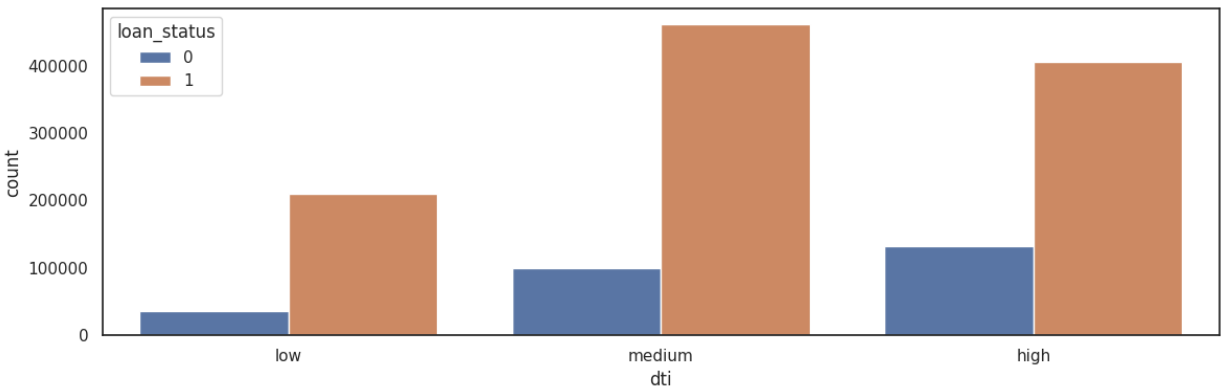

Figure 12: dti vs. loan status

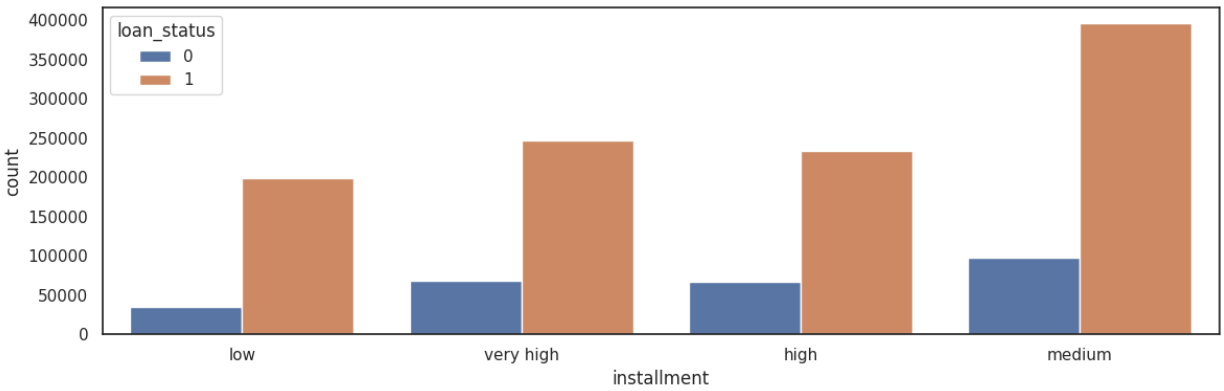

Figure 13: installment vs. loan status

ability of default for each applicant. You should thoroughly describe how you developed and validated your model and explain any assumptions you made.

The following steps have been followed in building the Binary classifier.

i. All the insignificant categories were dropped and the data was further cleaned to remove NaN's and droped all the columns with more than 50% missing values.

ii. The dummy variables were created for categorical variables whereever it was required.

iii. The dataset was split in to 80% training set and 20% Testing set.

iv. Normalized the categories with sklearn RobustScaler() to handle the outliers

v. Five binary classifiers have been modeld namely, LinearSVC, LogisticRegression, GaussianNB(), RandomForestClassifier, GradientBoostingClassifier, and XGBClassifier

vi. Also evaluated ANN for the binary classification

As the class imbalance was observed ie. more number of " Fully Paid" classes than the "Charged Off" Classes, the following approaches have been employed for training and evaluation of the model.

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

The accuracy metrics is good enough for a well-balanced class but not ideal for the imbalanced class problem.

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

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

For an imbalanced class dataset F1 score is a more appropriate metric. It represents the harmonic mean of precision and recall.

b. Resampling (Oversampling)

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

c.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.

The Evaluation Results are as follows

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

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

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

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

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

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

Figure 14: Classification Metrics without application of Data Imbalance handling Techniques

Evalution metrics with out application of imbalance techniques.

Achieved good accuracy with Naive Bayes classifier without employing any data imbalance handling techniques as shown in Figure 14.

Achieved good accuracy with Random Forest classifier by employing oversampling of minority class technique as shown in Figure 15.
The performance of all other classifiers was also improved considerably.

Achieved good accuracy with Random Forest and XGBoost classifiesr by employing SMOTE technique as shown in Figure 15.
The performance of all other classifiers was also improved considerably.

c. Imagine your client is considering entering the lending market but is very risk averse (they prefer low default rates even if it means accepting lower rates of return). Develop a strategy for entering this market.

Some things to consider:

i. Which locations should we target?

ii. To which segments of the population should we advertise?

iii. Any other helpful strategies you can think of to keep default rates low?

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

Figure 15: Classification Metrics with application of Data Imbalance handling Techniques-Resampling (Oversampling)

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

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

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

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

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

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

Figure 16: Classification Metrics with application of Data Imbalance handling Techniques(SMOTE)

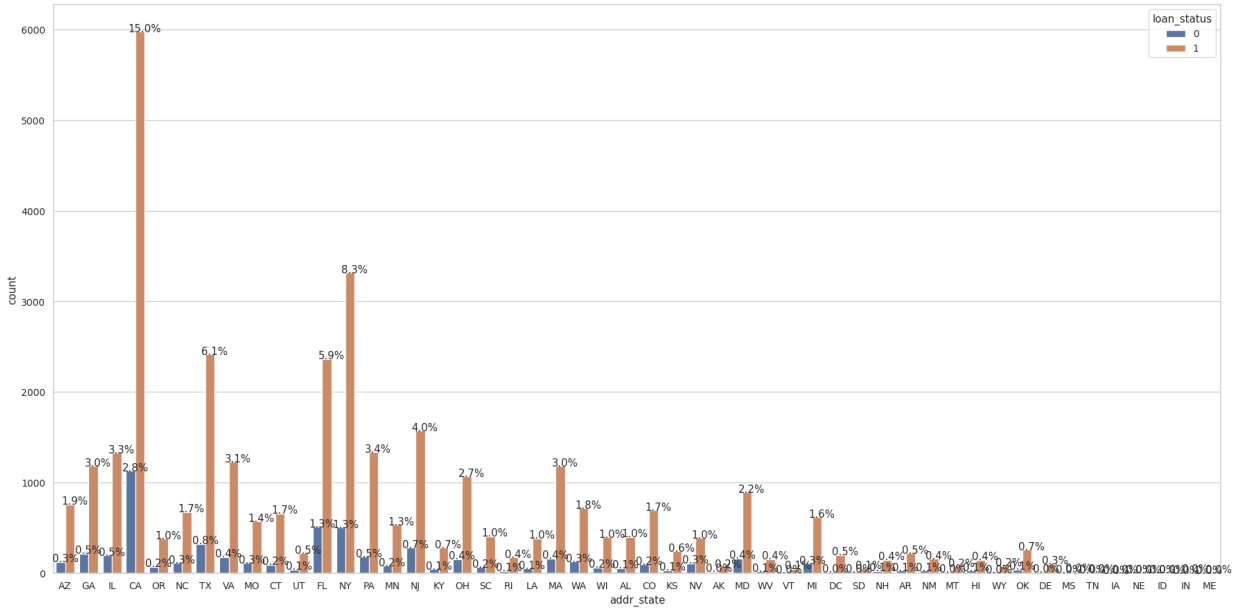

Figure 17: Loan default percentage accross states

The CA, NY, TX states had high number of applications and high default rate. The plot is represented in Figure 16.

Better to avoid small business loans followed by educational loans.The plot is represented in Figure 17.

The annual income, the loan duration or term, lower dti, the loan installment amount are the major factors to be considered before the new loan approvals.

The verification process should be revamped with more smarter processes as the verified customers deafulted quiet often.

:

3

Show us what you can do! This is an **optional free-form** section. Do you have a data science skill that can separate you from other candidates? This is your chance to show it off. Here are some ideas to get you started, but you can answer this section any way you'd like.
a. **Anomaly detection.** Can you spot any interesting anomalies or outliers in the data?

The distribution of the numerical continous data showed the skewedness and the presence of outliers as shown in Figure 18.

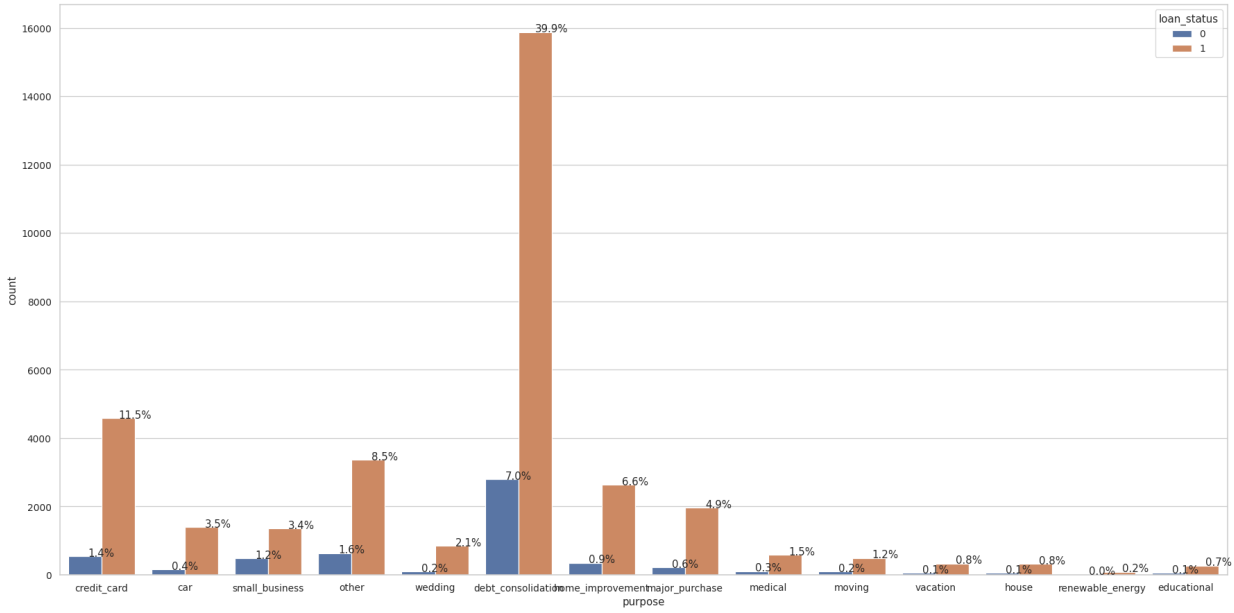

Figure 18: Loan default percentage accross differente purpose or segment

Sklearn's Robust Scaler has been used to correct the outliers.

b.**Visualization.** Create a visualization that reveals something interesting in the data or describes the data in a compelling way.

The correlation of differentnumerical features with loan status as shown in Figure 19.

c. **Kaggle Grand Master-level model performance.** Do you relish squeezing every last thousandth out of your F1 score? Develop a highly tuned, highly accurate model. Be sure to explain your approach to tuning and evaluating your model.

The F1-score of the model has been improved with robust scaling and with appropriate data imbalnce handling techniques.

d. **Data augmentation.** Can you find other public data sets that may improve your model? Ingest the data, join it to the original data set and explain how the new data did or did not improve your model's performance.

The provisioned data has got loan records from 2007 to 2011 with 42k applications. Further, the data has been augmented with the open source lending club data from kaggle website.

The new data has got the loans from 2007 to 2018 with approximately 2.2 million applications.

The clasification metrics without any data imbalance techniques is represented in Figure 20.

The clasification metrics with resampling imbalance techniques is represented in Figure 21.

The clasification metrics with SMOTE imbalance techniques is represented in Figure 22.

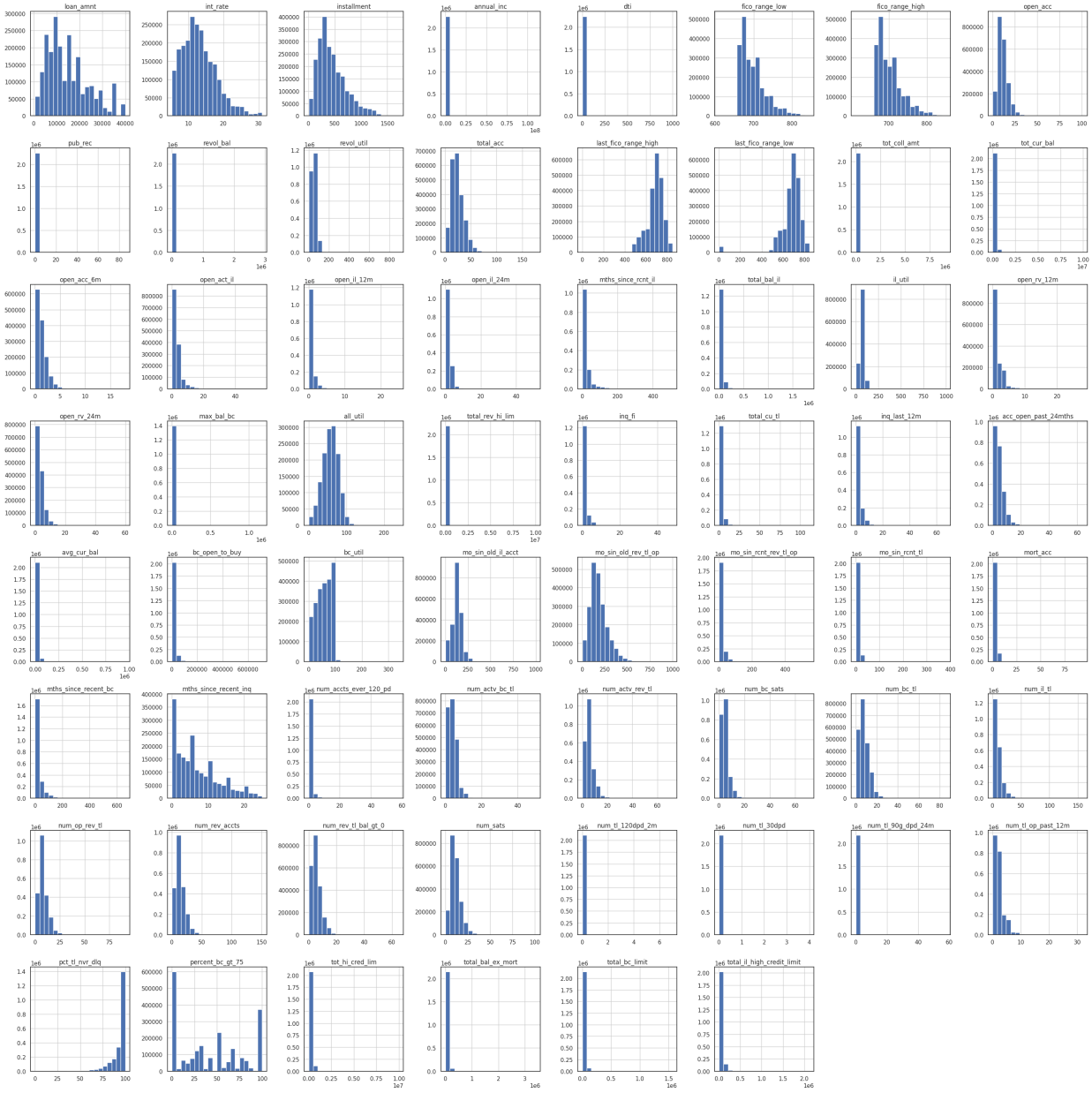

Figure 19: The data skewness and outliers

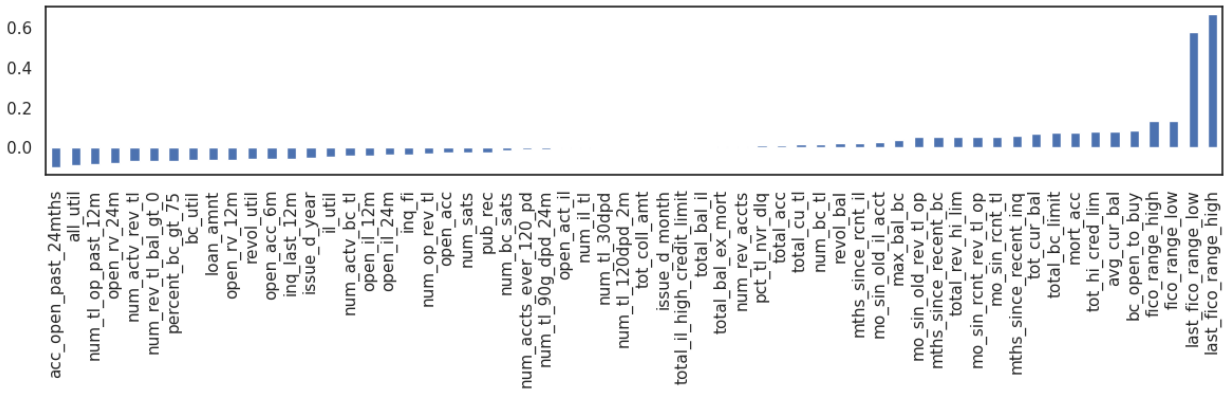

Figure 20: The data skewness and outliers

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

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

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

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

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

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

Figure 21: New Dataset-Classification Metrics without application of Data Imbalance handling Techniques

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

Figure 22: New Dataset-Classification Metrics with application of Data Imbalance handling Techniques-Resampling (Oversampling)

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

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

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

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

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

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

Figure 23: Classification Metrics with application of Data Imbalance handling Techniques(SMOTE)

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

Figure 24: ANN Classification comparison AT & T data and New LC Data

The clasification metrics with ANN classification with original data and imbalance data is represented in Figure 23.

e. **Be creative.** Come up with something we didn't think of and impress us with your amazing findings!

The net profits by asset class and default status is as shown in Figure 24.

The important features identified the model is as shown in Figure 25.

4 References

1. Link `https://www.kaggle.com/datasets/wordsforthewise/lending-club`
2. Link `https://pages.databricks.com/201902-EB-Loan-Risk-Analysis-with-Databricks-XGBoost.html`
3. Link `https://scikit-learn.org/stable/modules/model_evaluation.html`
All the links accessed on 28-03-2023

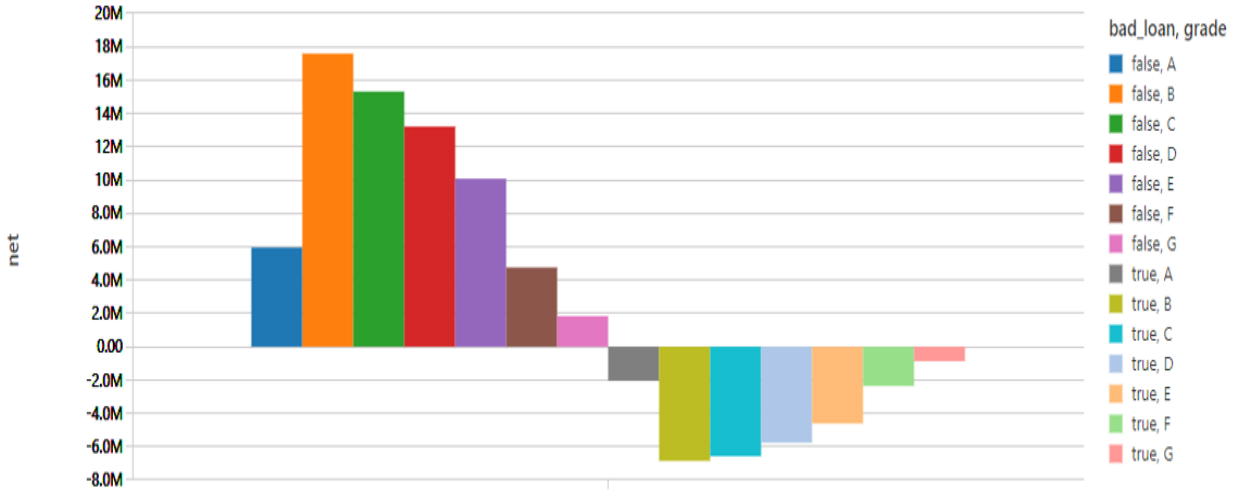

Figure 25: The net profits by asset class and default status

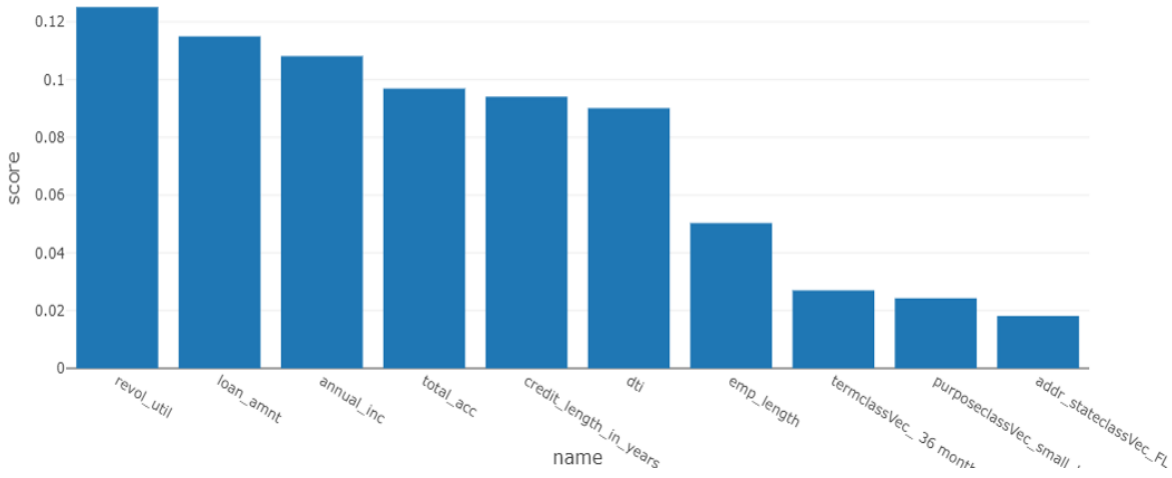

Figure 26: The important features identified the model