

**LENDING CLUB LOAN DATASETS**

**Post Graduate Program in Data Science Engineering**

*Location:* **Hyderabad** *Batch:* **DSE-FT JAN-21 HYD**

*Submitted By*

**GUNDLURU VENKAT RITISH**

**HARSHITHA MOR**

**MADANU VINITH XAVIER**

**VEENADHARI BEERAVELLI**

**R. SUMITH KUMAR**

**Under the Esteemed Guidance of**

**SRIKAR MUPPIDI**

**TABLE OF CONTENTS**

**CONTENTS…………………………………………...……….……………………………**

1. ABSTRACT
2. INTRODUCTION

2.1 DOMAIN AND FEATURE REVIEW

2.2 DATASET INFORMATION

2.3 PROBLEM STATEMENT

2.4 TARGET VARIABLE

3. EXPLORATORY DATA ANALYSIS

3.1 UNIVARIATE ANALYSIS

3.3 INSIGHTS FROM EDA

4. DATA CLEANING

4.1 DROPPING OF COLUMNS

4.2 MISSING VALUE TREATMENT

4.3 TRANSFORMATION

5. BIVARIATE ANALYSIS

6. UNI FEATURE SELECTION USING STATISTICAL TEST

7. MACHINE LEARNING MODEL

**8. FEATURE SELECTION**

8.1 P-VALUE METHOD

9. FUTURE WORK

**1.ABSTRACT**

The financial tech industry generally deals with a lot of preprocessing requirements of loans before they are analysed by lenders, investors and bankers. What’s usually a manual process in banks and similar financial institutions is automated utilizing highly active, adaptable machine learning models. These models make use of previously recorded financial data to learn the patterns in loan consumption, approval and defaulting – and predict the security of providing a borrower with the loan amount they have requested. This is one of the many utilities for which data science is used in the field of finance.

In this project we’re taking the historical financial data of Lending Club, a peer-to-peer loan lending institution which allows direct investment across 28 states in the United States of America. The data extracted is from the final quarter of 2018. The goal is to analyse and interpret our data to build a machine learning model based on binary classification that predicts the defaulter status of a future borrower applicant at Lending Club.

The blurb of this report includes the summary of our problem statement, data overview, feature analysis, exploratory data analysis, data cleaning, statistical analysis of all the features in the current dataset and the base classification model involving the same.

**2. INTRODUCTION**

**2.1 DOMAIN AND FEATURE REVIEW**

Our problem statement currently involves the finance domain, in particular financial technology. Which according to an article written by Julia Kagan (Financial Technology – Fintech, Investopedia, August 28, 2020) is defined as follows:

“*Financial technology (Fintech) is used to describe new tech that seeks to improve and automate the delivery and use of financial services. At its core, fintech is utilized to help companies, business owners and consumers better manage their financial operations, processes, and lives by utilizing specialized software and algorithms that are used on computers and, increasingly, smartphones. Fintech, the word, is a combination of "financial technology.””*

Financial technology is embedded into our everyday process chains starting right from the financial service providers to the average consumer – ranging from business models that help institutions manage their finances better to mobile wallets that make it easy for consumers to access their own monetary resources at the click of a finger.

The application of this domain in our current project relates to the loan/credit risk analysis that borrowers’ information undergoes before their loans are either denied or accepted based on the risk they pose of being a loan defaulter/delinquent. The manual process of reviewing a loan application is prone to human errors and/or personal bias. It’s also extremely time consuming. Automating this review process by boiling down our applicant history into a set of variables to be analysed by a machine learning model will reduce the human error and avoid bias. It also reduces the time required to go through an applicant’s history.

An applicant’s history that is required by the financial institution for loan approval include their income portfolio, credit history, past delinquencies, current loan repayment status, past lending history, home ownership of the borrower etc.

This history is collected in the form of different features that are collated into datasets for the sake of model building. They include financial account history, number of delinquencies, time since the borrower’s last delinquency, since their last account has been opened (be it a credit account, revolving trade or otherwise), their annual income, debt-to-income ratio, amount owed without considering one’s mortgage, employee title, work experience etc.

Since machine learning models are dealing with such sensitive data it is important that they are built to be as robust and adaptable as possible.

**2.2 DATASET INFORMATION**

Lending club is one of the first peer-to-peer fintech loan lending financial institutions. It was started in the year 2007 and has upto 3 million members in the club till date including both borrowers and investors. They also provide a range of financial services and products like a regular bank does. According to Lending Club’s website their goal involves:

“*Helping Americans meet their life goals:*

*Since 2007, more than 3 million members have joined the Club to help reach their financial goals. As the only full-spectrum fintech marketplace bank at scale, our members can gain access to a broad range of financial products and services through a technology-driven platform, designed to help them pay less when borrowing and earn more when saving.”*

The data set contains the profile of various borrowers between the time period of June 2007 to December 2018, holding information on the number of accounts (if delinquent or not), balance to credit limit, average current balance of all accounts, number of charge- offs etc.

## SELECTION OF DATA SAMPLE:

The below mentioned methods would be analysed in order to obtain better results and have a realistic model prediction, without deviating from the business goal:

The following three data sampling methods have been selected for the particular dataset sampled from the original dataset used in this project:

1. Stratified sampling: we will take a split of whole dataset which might be around 97.5% and 2.5% as the population size is around 22lakhs and we will try to get equal no.of records from each year (of the 11years data we have)

2. Thus each year has different external factors affecting so keeping our study of concentration only for four months or 3months of a particular year would be more appropriate considering even the business problem statement

3. Or we can also take the two different sample by divinding the dataset by the independent variable "term" ( 36 months and 60 months ) and study the results accordingly.

We’ve utilized the second approach for our project, for the sake of most recent data extraction and time sensitivity.

**SIZE OF DATA SAMPLE CHOSEN:**

|  |  |
| --- | --- |
| Number of Attributes | 145 |
| Number of records | 150792 |

This sample of data is being utilized for better results because of the following reasons:

* Financial trends are time sensitive. What’s true for the first quarter of a year does not hold true for the next quarter. Building a model that learns from years worth of financial history will only destabilise the model since there are too many trends being captured in the data
* It’s also imperative to note that lack of autoML tools for the purpose of model building in our data can lead to the reduction in the data holding capacity for regular ML tools like python, SQL, tableau etc. Especially since the dataset can get larger when the requirement for feature engineering sets in.

**2.3 PROBLEM STATEMENT**

The divergence of the two entities which are investing and borrowing provided by the company has suitably its dependence on the availability of prominent information of each individual. The following characteristics have been taken into consideration for estimating every account:

* 1. Status of the property: if it is rented, owned, or is under mortgage
  2. Last date of delinquency, if any: recent failure in paying the due on time
  3. Job profile and duration
  4. Loan amount that exists already, if taken previously

Lending Club provides an online platform to both lenders and investors to cater the needs. Lenders such as the banks or any individual investors receive an opportunity in lending money at low interest rates while the investors have an opportunity to earn money out of the interest acquired while the loan is being served. Lending Club gives institutions of all kinds, the ability to invest at scale. The banks being one of the prime investors, are enabled in purchasing a loan and also given provision to offer Lending club products to their customers.

The target variable is loan\_status and according to the problem statement, we categorize it into fully paid and a defaulter. This is a binary classification problem and based on the insights from the model building process it would help the bank segregate people and thus making their job easy and profitable.

**2.4. TARGET VARIABLE**

**loan\_status**: Current status of the loan (in terms of repayment)

Statuses include:

* Current: Loan is up to date on all outstanding payments.( loan\_status = 0)
* In Grace Period: Loan is past due but within the 15-day grace period. ( loan\_status = 1)
* Late (16-30): Loan has not been current for 16 to 30 days. ( loan\_status = 1)
* Late (31-120): Loan has not been current for 31 to 120 days. ( loan\_status = 1)
* Fully paid: Loan has been fully repaid, either at the expiration of the 3- or 5-year year term or as a result of a prepayment. ( loan\_status = 0)
* Default: Loan has not been current for an extended period of time. ( loan\_status = 1)
* Charged Off: Loan for which there is no longer a reasonable expectation of further payments. Upon Charge Off, the remaining principal balance of the Note is deducted from the account balance.( loan\_status = 1)

Since taking any form of risk when it comes to financial aid results in loss, only two of the above statuses have been assigned the binary value 0, indicating non-defaulters of loan, and the rest of the status have been assigned 1, indicating defaulters of loan. This is because even grace periods are delinquencies to begin with. A day’s delay in full loan repayment is still a record that doesn’t leave one’s credit report seven years after it has been recorded – and for good reason.

**3. EXPLORATORY DATA ANALYSIS**

**3.1 UNIVARIATE ANALYSIS**

We have done uni-variate analysis for all the columns present in the original dataset (145-columns).

The process of the analysis undertaken is listed as follows:

1. Analyzed the data using basic statistics from the describe function of the columns
2. Visualizing the columns using various plots like box plot (for checking outliers)

and distribution plot (for checking the spread of the data) for numerical columns

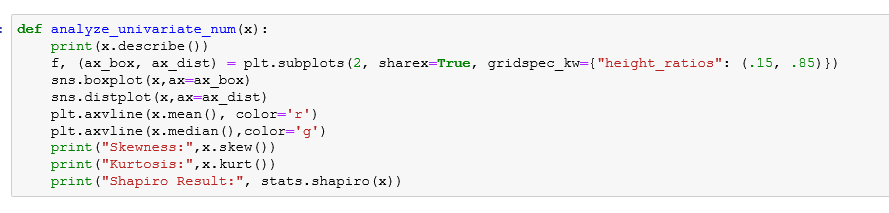
and countplot (for checking the variance among variables in a column) for the categorical columns

1. Checking for the normality of data using skewness and kurtosis

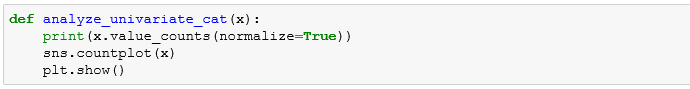
The results from uni-variate analysis were used for dropping initial columns with null values and missing value treatment.

For this we have written two user defined functions to analyse numerical and categorical variables respectively.

**For numerical variables:**

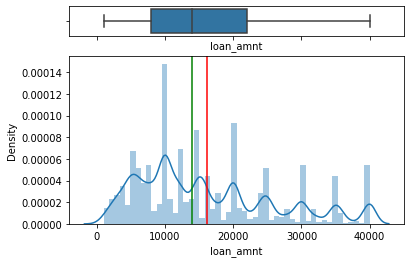


**For categorical variables:**



Lets take few variables to understand the above functions.

1. loan\_amnt(numerical): Below are metrics for skewness and kurtosis along with numerical data description of loan\_amnt variable



count 150792.000000

mean 16095.824712

std 10179.502959

min 1000.000000

25% 8000.000000

50% 14000.000000

75% 22000.000000

max 40000.000000

Skewness: 0.7519410576240981

Kurtosis: -0.3353669255517770

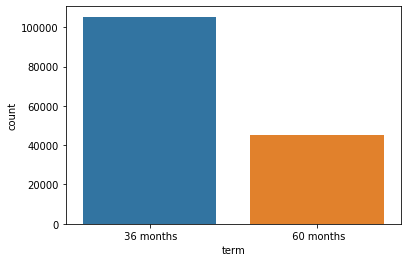
Conclusion: From the boxplot we can say that there are no outliers in the loan\_amnt.There are multiple peaks so we can do transformation. The data also has to be scaled.

Consensus: Loan amount is the amount of loan provided to the borrower on loan approval by the investor via a financial corporation. At times the investor and corporation are the same as seen in the case of banks. For LendingClub this amount ranges anywhere from $1000 to $40,000 USD. It’s imperative to note that while a borrower can request for a loan amount he wants, it can be altered further down the line based on his financial history. This variable is important because it helps to gauge how much debt the person can handle. From the above graph it is clear that borrowers are lending club mostly borrow within the range of 1000 USD to 15000 USD.

1. term(categorical): Below are the percentage metrics and countplot of term

36 months 0.69887

60 months 0.30113

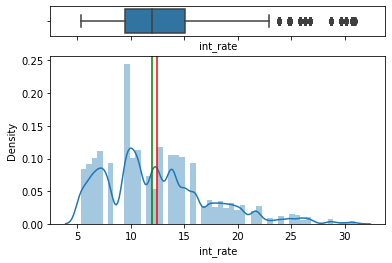


Conclusion:There is good percentage of values for both 36months and 60months classes. The classes are well balanced across our dataset. They could be turned to dummies for model building.

Consensus: Term indicates the timeline of a borrower’s loan repayment plan. For the most part it is indicated in terms of months rather than years or weeks. This is because interest rate is applied on a monthly basis for every installment that is paid. At LendingClub every repayment plan spans across a term of either 36 months or 60 months – based on their financial stability. For example, if a person was assigned a term of 36 months, they can pay the full amount before the term ends without any pre-payment penalties but the same doesn’t apply if repayment was delayed past the term span. And delinquencies on any monthly installment incur a penalty on the borrower past the 7-day grace period.

1. int\_rate(numerical): Below are metrics for skewness and kurtosis along with numerical data description of int\_rate variable

count 150792.000000



mean 12.455130

std 5.028557

min 5.310000

25% 9.430000

50% 11.980000

75% 15.050000

max 30.990000

Skewness: 0.8606157071452561

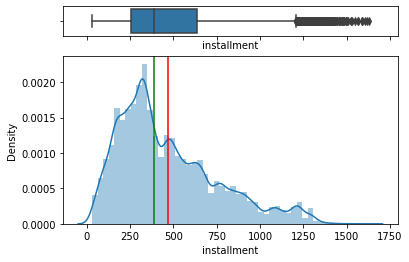
Kurtosis: 0.5741871241572105

Conclusion: There’s heavy presence of outliers that can be treated using transformation. Data is not normal. Data is moderately positively skewed . Data follows a leptokurtic distribution

Consensus: int\_rate is the interest rate assigned by LendingClub once the borrower has been assigned a grade and a subgrade based several factors, such as credit history and rating, the amount one wants to borrow, and their total debt-to-income ratio. The rating/subgrade assigned to each borrower based on their credentials doesn’t just depend on borrower information but also the year and quarter of the year the loan is being lent. Interest rates can fluctuate once the repayment term is over. LendingClub offers a rate checking facility on their website which allows borrowers to know how much interest they will be charged. Here in our data most borrowers are charged between 5-15% per month.

1. installment(numerical): Below are metrics for skewness and kurtosis along with numerical data description of installment variable.

count 150792.000000



mean 468.076601

std 288.642302

min 29.760000

25% 254.000000

50% 386.840000

75% 635.660000

max 1628.080000

Name: installment, dtype: float64

Skewness: 0.9283743045688089

Kurtosis: 0.27568191450243207

Conclusion:  The Upper bound has to be increased inorder to include all the outliers(they are very densely populated/close to each other). Data is not normal .Positively skewed.Follows leptokurtic Distribution. Utilisation of transformation to normalize data required

Consensus: installment is the monthly repayment amount a borrower is assigned after loan disbursement. It depends on the term chosen for repayment and financial stability of the borrower. This feature in particular indicates the monthly installment amount irrespective of interest rate. Interest is charged on a monthly basis and fluctuates across a financial year, so person’s total monthly payment would be interest rate + fixed monthly installment. Two borrowers borrowing the same amount can have different terms and hence different fixed installment amounts to be paid. Any payment towards the loan consolidation should never go below the total monthly payment amount(considering the interest)

1. sub\_grade(categorical): Below are the percentage metrics and countplot of sub\_grade

A2 0.046627

A1 0.042787

D1 0.032343

D2 0.030804

D3 0.029299

D4 0.026109

D5 0.024438

E5 0.008542

E3 0.007938

E4 0.006784

E2 0.006400

E1 0.004377

F1 0.003840

G1 0.001021

F2 0.000749

F3 0.000736

F5 0.000584

F4 0.000564

G2 0.000060

G4 0.000040

G3 0.000033

G5 0.000013

B5 0.065023

B1 0.063538

B2 0.062231

C1 0.060812

B4 0.057888

A4 0.057304

C2 0.053816

C3 0.053816

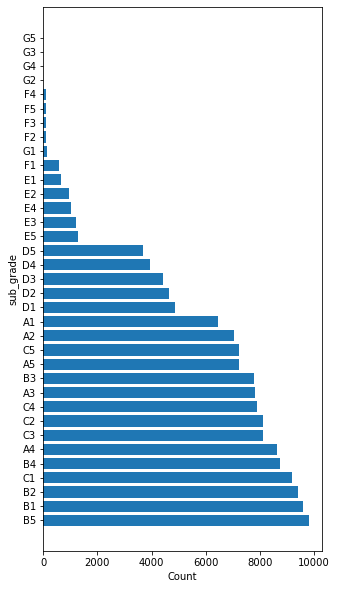
C4 0.052271

A3 0.051860

B3 0.051561

A5 0.047967

C5 0.047827



Conclusion: Our subgrades range from A1 to G5. In our current dataset the highest number of borrowers were assigned the B5 grade. And the least number of borrowers were given the G5 grade (the lowest grade possible)

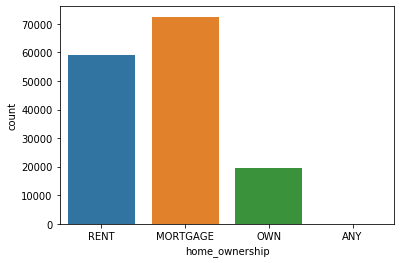
Consensus: Every borrower that applies for loan at lending club is given a grade(A to G) and further a sub grade(1 to 5) amidst those grades to rate a borrower based on their credit risk. The higher the grade the lower the credit risk of the person. And based on the grade assigned to the person they are charged a certain Annual Percentage Rate (interest rate + origination fee + true cost of the loan borrow). The final interest rate for each loan grade is the result of the following equation:

Lending Club Base Rate + Adjustment for Risk & Volatility

The Adjustment for Risk & Volatility is designed to cover expected losses and provide higher risk-adjusted returns for each loan grade increment from A1 to G5.

Sub grade is a highly defining factor about risk that comes with any particular borrower, given that it’s assigned after evaluating the borrower’s financial profile thoroughly.

1. home\_ownership(categorical): Below are the percentage metrics and countplot of home\_ownership.



MORTGAGE 0.480443

RENT 0.390982

OWN 0.128548

ANY 0.000027

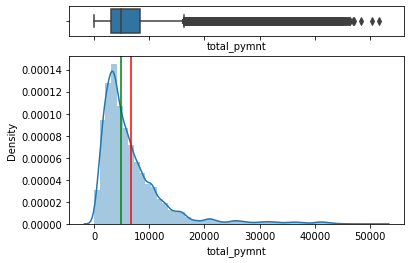
Name: home\_ownership, dtype: float64

Conclusion: The only feature engineering activity to be done on this variable is to create dummy variables.

Consensus: Home\_ownership feature in our dataset indicates the status of home ownership of our borrower. Whether it be that they own a home, or currently own one with mortgage payments going on or if they live in a rented space. There are also other kinds of residencies wherein a tenant lives in a space lease-free which might be possible if the housing costs are being paid of by their employer or if they are staying with a parent/friend who doesn’t charge them. Home ownership status is important in loan approval because it serves as proof of the person’s financial stability or pre-existing debt weight (if they are paying mortgage) – and also as collateral in the event of the borrower turning out to be a delinquent/defaulter. In the USA a lot of citizens prefer owning a home and it can be seen in our data too. A lot of people prefer to take up mortgage irrespective of the debt just to own a home.

1. total\_pymnt(numerical): Below are metrics for skewness and kurtosis along with numerical data description of total\_pymnt variable

count 150792.000000



mean 6779.175168

std 6289.429453

min 0.000000

25% 3023.770000

50% 4958.780000

75% 8360.750000

max 51653.389338

Name: total\_pymnt, dtype: float64

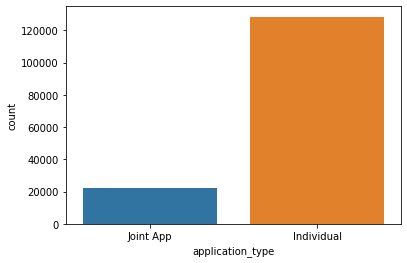
Skewness: 2.735106894842325

Kurtosis: 9.659036267009078

Conclusion: It is heavily leptokurtic and positively skewed. It requires transformation to treat the heavy presence of outliers and the skewness of the data. Capping might also be required since our outliers are so densely populated together near the upper limit of our data.

Consensus: total\_pyment is the total amount paid off by the loan borrower till data out of the full amount funded to them. Total amount should usually be equal to the funded amount by the time the applicant’s term reaches completion in order to not be a defaulter of loan. It’s usually not the case if the borrower has defaulted on the loan or has been charged-off. Before term completion though, this amount needs to be equal to the product of number of months and monthly installment – indicating the person has been up to date with all payments without any delinquency. It can be higher than that too if a person decided to pre-pay a large amount in lesser time to avoid the interest rate for all the months included in their term plan.

1. application\_type(categorical): Below are the percentage metrics and countplot of application\_type



Individual 0.851192

Joint App 0.148808

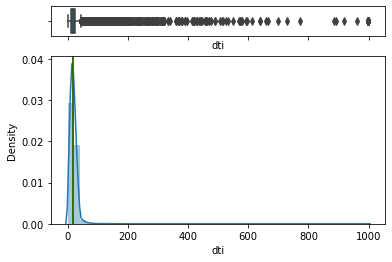
Name: application\_type, dtype: float64

Conclusion: Most borrowers are individual applicants rather than joint borrowers. The next step in utilizing this variable would be to get dummies for it for our machine learning model

Consensus: Application type indicates if a person has individually applied for a loan or if the loan was borrowed by two people under a joint application. Usually joint applications can be a good thing because it ensures more stability and security in loan repayment. There are always deviations wherein neither of the applicants have stable income sources or one of them has a lot of prior debt, and it can also be disadvantageous if there were to be asset dispute amongst the co-applicants in the future. Both applicants are charged equally, wherein the loan payment amount is split amongst them thus providing tax relief for the same amount an individual borrower applies for. The most popular reason for joint loan applications if for home loans amongst married couple, closely followed by business partners who are starting something new. For a range of reason joint applications are good as long as the co-applicants work well together in handling the loan. There’s lesser chance for them to default a loan.

1. dti(numerical): Below are metrics for skewness and kurtosis along with numerical data description of dti (Debt-to-income ratio) variable.

count 150429.000000



mean 19.616939

std 21.248328

min 0.000000

25% 11.230000

50% 17.650000

75% 25.030000

max 999.000000

Name: dti, dtype: float64

Skewness: 24.304971844626493

Kurtosis: 960.4529148714875

Conclusion: Our data is heavily right skewed. Transformation and capping is required to treat the heavy presence of outliers.The distribution is leptokurtic in nature

Consensus: Debt-to-income ratio of a borrower is highly valuable in gauging the credit risk a certain borrower poses. Because it is the ratio between the total monthly debt payments and the monthly income of a person and gives insight on how much additional debt a person could handle. While most financial institution consider mortgage in the calculation of DTI, LendingClub excludes mortgage payments and the requested loan from LC for more stable and reliable numbers. The lower the DTI the better the borrower’s chances of getting their loan approved. A DTI below 28% is considered a really good DTI and this limit can sometimes be stretched to 37% but that’s still pushing the default risk. At LendingClub, 40% and below is considered ideal for borrowers before they send in their loan application. There are various ways to mitigate high DTI, primarily an increase in monthly income and a huge amount of debt consolidation.

10. annual\_inc(numerical): Below are metrics for skewness and kurtosis along with numerical data description of annual\_inc (annual income) variable.

50% 6.500000e+04

75% 9.500000e+04

max 9.930475e+06

Name: annual\_inc, dtype: float64

Skewness: 48.017853354712095

Kurtosis: 4745.784693082272

Shapiro Result: (0.38317781686782837, 0.0)

count 1.507920e+05

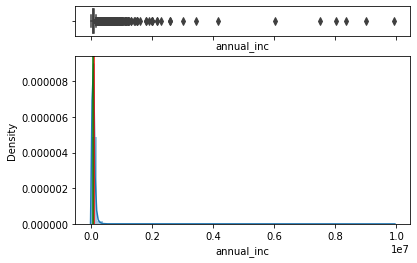
mean 7.858652e+04

std 8.013942e+04

min 0.000000e+00

25% 4.500000e+04

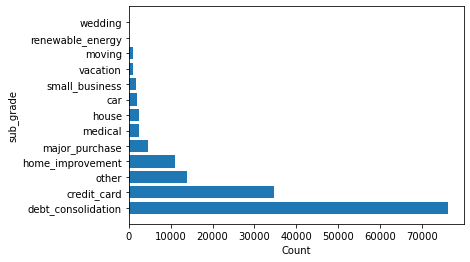
Conclusion: 1.Our data is heavily right skewed. 2. Heavy presence of outliers that can be treated with transformation. 3. The distribution is leptokurtic in nature 4. The outliers in our data are very huge compared to the mean and median of the spread



Consensus: Annual income is the annual earnings of a borrower and this case this feature indicates the annual income of the primary applicant if the loan was a joint application. Annual income is another heavily applicable variable that’s considered for the sake of loan approval or rejection. Because a high income indicates good ability to handle debt/additional debt. It indicates low credit risk. If a person doesn’t have high annual income they can still go through loan approval either by making huge down payments at the get go, adding on a co-applicant(although this is risky), or by paying off a huge amount of debt before applying for any amount of loan (a tactic used to decrease our DTI).

11.Purpose(categorical): Below are the percentage metrics and countplot of purpose

debt\_consolidation 0.505173



credit\_card 0.229873

other 0.091417

home\_improvement 0.072922

major\_purchase 0.029730

medical 0.016327

house 0.015684

car 0.012434

small\_business 0.011784

vacation 0.007229

moving 0.006831

renewable\_energy 0.000564

wedding 0.000033

Name: purpose, dtype: float64

Conclusion: The most frequent purpose for lending a loan at Lending Club turns out to be debt consolidation which is paying off previous debts that are nearing their term completion. It can because of student loans, business loans, car loans, credit delinquencies that need to be paid off etc.

Consensus: Purpose of a loan dictates the reason a person wants to lend a loan from an institution. Lending Club provides loan under personal loan disbursement. While credit history, income, previous debts all matter the most when it comes to evaluating loan risk, purpose dictates the interest rate that an institution charges the applicant to an extent. For example certain institutions/banks charge less interest for credit card debt but more for home improvement purpose loans. At Lending Club, this variable ties into the sub\_grade assignment of a person so that they can be charged the appropriate interest based on their purpose amongst numerous other factors.

**EDA insights for the remaining features**

**FINANCIAL ACCOUNTS BASED FEATURES:**

These are all the features that contain data of all financial accounts owned by the applicant, including but not limited to number of accounts and timelines of the same with respect to open trades, credit lines, revolving credit, bank card accounts, installment accounts, mortgage and number of satisfactory bank accounts. They are as listed below:

*open\_acc, total\_acc, earliest\_cr\_line,open\_acc\_6m,open\_act\_il,open\_il\_12m,open\_il\_24m, open\_rv\_12m,open\_rv\_24m, total\_cu\_tl, acc\_open\_past\_24mths,mo\_sin\_old\_rev\_tl\_op, mo\_sin\_rcnt\_rev\_tl\_op, mo\_sin\_rcnt\_tl, mort\_acc, num\_accts\_ever\_120\_pd, num\_actv\_bc\_tl, num\_actv\_rev\_tl, num\_bc\_sats, num\_bc\_tl,num\_il\_tl, num\_op\_rev\_tl, num\_rev\_accts, num\_rev\_tl\_bal\_gt\_0, num\_sats, num\_tl\_op\_past\_12m, mths\_since\_rcnt\_il, mo\_sin\_old\_il\_acct, mths\_since\_recent\_bc.*

Conclusion: The data for all these variables is heavily right skewed and in need of transformation for the treatment of outliers. Most of these features share similar distributions with other variables in this list, similar down to the skewness metric derived. Variables *mo\_sin\_old\_il\_acct* and *mo\_sin\_old\_rev\_tl\_op* are more normal than the rest. [Because they indicate the number of months since the borrower’s oldest revolving trade and installment accounts have been opened and can be a large timeline depending on when the accounts have been opened. The oldest installment account in our data dates back to 1950]

Consensus: The features involve data about the oldest and most recent financial accounts opened with respect to open trades, credit lines, revolving credit, bank card accounts and installment accounts, and the months since then and the number of accounts opened since their oldest account.

While the number of accounts in itself don’t affect a person’s credit report, the balance remaining in them along with any debt owed to any of these accounts do impact a person’s loan approval status. The higher the number of accounts the more chances for higher debt consolidation, and hence more loan default risk. But it’s not always the case. The impact these accounts can have on loan approval depends on whether or not the person is timely on their payments, with little to no delinquencies.

**BALANCES AND CREDIT LIMIT BASED DATA/FEATURES:**

The features discussed in this section are features related to the balances, limits and credit percentages of the financial accounts discussed above. They are as listed below:

*revol\_bal, tot\_cur\_bal, tot\_coll\_amt, total\_bal\_il, total\_rev\_hi\_lim, tot\_hi\_cred\_lim, total\_bal\_ex\_mort, total\_bc\_limit, total\_il\_high\_credit\_limit, il\_util, bc\_util, percent\_bc\_gt\_75, bc\_open\_to\_buy, revol\_util, all\_util*

Conclusion: The data for all these variables is heavily positively skewed and in need of transformation for the sake of outlier treatment. They are all leptokurtic in nature of distribution. Variables *bc\_util* and *revol\_util* are less skewed than the rest in this bunch (this is because the values in this data are not as huge in magnitude as the other features in this section. They indicate ratios of balance to limits of revolving and bankcard accounts of the borrower.)

Consensus: The features involve the data regarding the total credit limit, total revolving account limits, balances of credit and revolving accounts – taking into consideration all accounts owned by the borrower. As stated in the previous section these balances are limits are important to gauge how much debt a person is paying off on an average. If the ratio of balance to limit is high it indicates that the borrower has been punctual and risk free with respect to the account debts they owe. If the ratio is really low, it indicates that the person has not been careful with their debts or were not in a position to do so – and hence can’t handle additional debt well.

**BANKRUPTCIES, PUBLIC RECORDS AND DELINQUENCIES:**

The below features indicate any public records, delinquencies or bankruptcy based information regarding the borrower: They are listed as below:

*delinq\_2yrs, pub\_rec, pub\_rec\_bankruptcies, pct\_tl\_nvr\_dlq*

These features are negatively impacting in nature to a person’s loan approval since public records show the borrower’s previous failures to repay debts, and hold up their financial status.

Conclusion: All the variables in this section are heavily positively skewed except for pct\_tl\_nvr\_dlq which is heavily negatively skewed. They all require transformations to treat the skewness and outliers. All of them share a leptokurtic distribution

Consensus: These features indicate all the public derogatory records that remain in a person’s credit report past the 7-year marks since they have occurred. It’s not a good sign to have them present in one’s loan application and have legal consequences if hidden while providing information to investors. Most of the borrowers in our data do not have any public records and hence why these variables are heavily populated with zero. Except for pct\_tl\_nvr\_dlq which indicates the number of trades that were never delinquent (which is logical since having 0 public records are delinquencies indicates high percentage of non-delinquent accounts)

**FINANCIAL INQUIRIES**

These features include data regarding any financial inquiries a borrower might have had regarding the personal loan they are about to acquire. They are as listed below:

*inq\_fi, inq\_last\_12m, mths\_since\_recent\_inq*

Conclusion: All the variables in this section are positively skewed, mths\_since\_recent\_inq less than the other two. They require transformation to treat outliers and skewness. The features in this section all follow leptokurtic distribution except for mths\_since\_recent\_inq which is more platykurtic.

Consensus: Usually while considering a person’s credit score/credit report there are general inquiries and there are hard inquiries. Our data has no distinction between the two. Credit reports are affected if a person has more than one hard inquiry in the months leading up to a person’s loan approval. Hard inquiries lower credit score by several points and will remain on the credit report two years after they have been attempted.

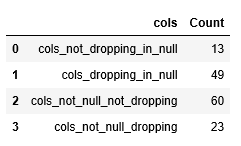
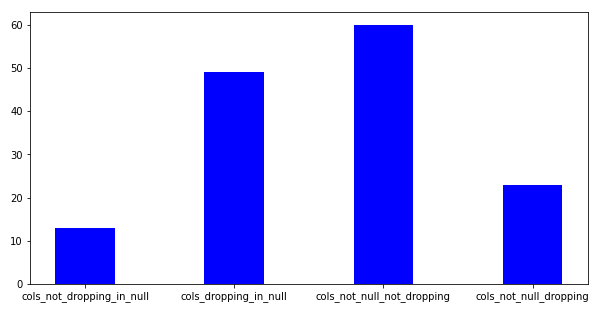
**4.Data Cleaning**

**4.1 Dropping of columns**

While performing the initial Data Cleaning process of the variables, below are the reasons that have been narrowed down for dropping the features:

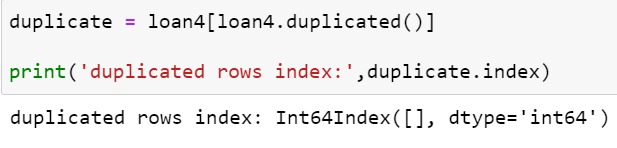
1. A very low variation among the classes of the variable (Eg: pymnt\_plan)
2. High percentage of missing values (Range Chosen: 85-100%) (Eg: num\_tl\_30dpd, hardship\_flag)
3. Columns with repetitive information (Eg: funded\_amnt)
4. Low contribution to target variable from a business perspective,(Eg:policy\_code)

Due to the huge number of attributes we have divided all the attributes into the following categories.

**4.2 Checking for Duplicate Records**

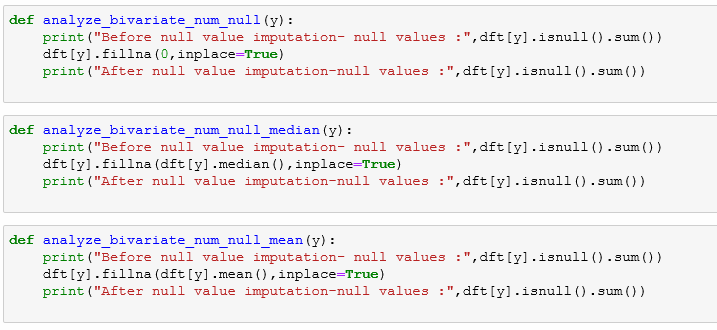
For such a huge data sample that we've acquired it's more than common to see duplicated records, wherein all values across all features are the same for multiple observations in our dataset. It's imperative to remove such records for a more accurate model. From the snippet shown below, it's evident there are no duplicated records in our dataset (the index list for duplicated records is an empty list)

****

**4.2 Missing Value Treatment**

From the above graph we can infer that there are 13 variables for which has to be treated for missing values. For the base model preparation we imputed the missing values based on the skewness of the data with mean, median(numerical) and mode(categorical). We have formulated the following user defined functions for the above mentioned imputation methods.

Here the central tendencies of our data distribution have been used for the sake of missing value imputation. Mean imputation was conducted on those variables which despite being skewed have a really low difference/distance between their mean and median. Median imputation has been utilised for variables with high skewness and a large difference between mean and median of their spread. Mode was used for imputing categorical features with missing values, where null values were replaced with the most frequent category/class in that particular feature.



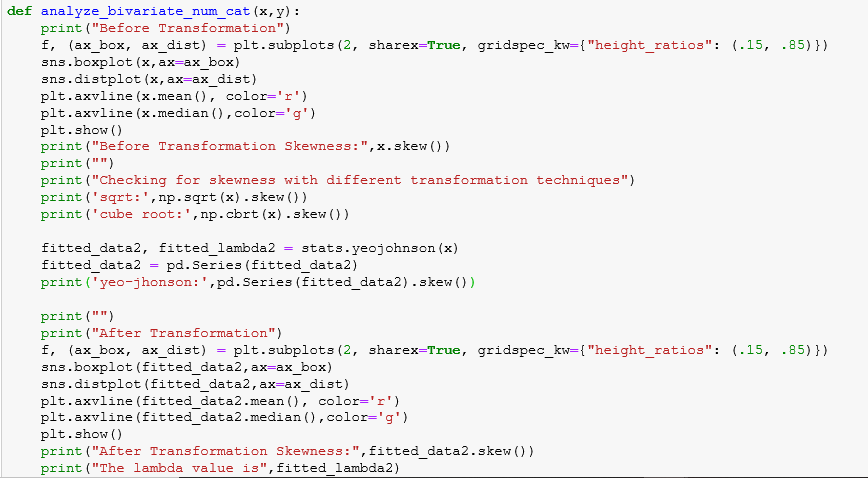
|  |  |  |
| --- | --- | --- |
| **Sr No** | **Column\_Name** | **NUll\_%** |
| 1 | il\_util | 16.61626611 |
| 2 | mths\_since\_recent\_inq | 12.42174651 |
| 3 | mths\_since\_rcnt\_il | 3.9033901 |
| 4 | mo\_sin\_old\_il\_acct | 3.9033901 |
| 5 | bc\_util | 1.427131413 |
| 6 | percent\_bc\_gt\_75 | 1.387341504 |
| 7 | bc\_open\_to\_buy | 1.384025678 |
| 8 | mths\_since\_recent\_bc | 1.300466868 |
| 9 | dti | 0.240728951 |
| 10 | revol\_util | 0.135285692 |
| 11 | all\_util | 0.026526606 |
| 12 | avg\_cur\_bal | 0.007294817 |
| 13 | pct\_tl\_nvr\_dlq | 0.000663165 |

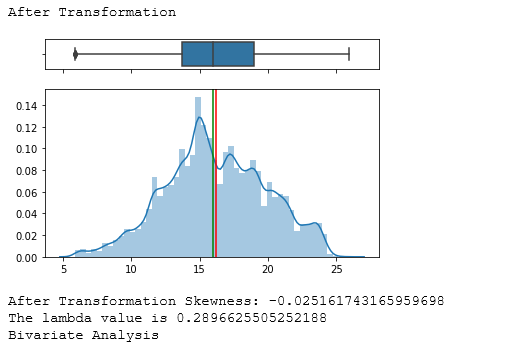
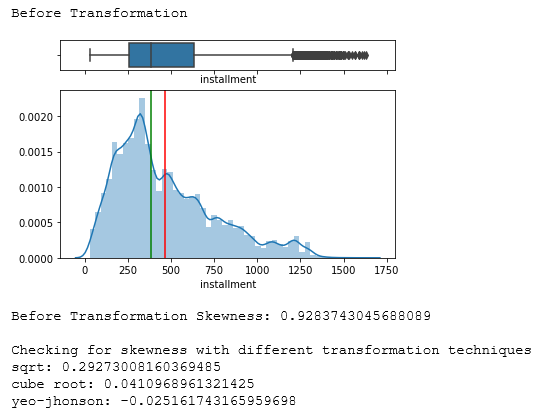
**4.3 DATA TRANSFORMATION**

After the univariate analysis of each feature, we could infer that dropping of columns and transformation was required for many variables to reduce high skewness and heavy outliers so as not to lose the data.

1. We have used square-root, cube-root, log and yeo-johnson transformation and evaluated skewness metric for the same
2. The transformation with the least skewed value was chosen from the above techniques and applied to our data.
3. The reduction in skewness and no.of outliers has been observed after transformation

For this we have written two user defined functions to analyse numerical and categorical variables respectively.





From the above graphs, we can infer that the skewness has been reduced considerably and the outliers have been handled accordingly

|  |  |  |  |
| --- | --- | --- | --- |
|  | Before transformation | After\_Transformation | Power transform parameter |
| loan\_amnt | 0.751941058 | -0.050876989 | 0.33080602 |
| int\_rate | 0.860615707 | 0.0031278 | -0.040903366 |
| installment | 0.928374305 | -0.025161743 | 0.289662551 |
| annual\_inc | 48.01785335 | 0.428105225 | 0.271010462 |
| delinq\_2yrs | 6.592775722 | 2.013992928 | -7.742160556 |
| open\_acc | 1.28774409 | 0.001849103 | 0.160732016 |
| pub\_rec | 14.19358187 | 2.098154354 | -9.701301396 |
| revol\_bal | 10.81329096 | 0.148140149 | 0.269135431 |
| total\_acc | 1.114742749 | -0.005984733 | 0.272993794 |
| out\_prncp | 0.791033351 | -0.694805906 | 0.353542081 |
| total\_pymnt | 2.735106895 | 0.070446347 | 0.131394711 |
| total\_rec\_prncp | 3.105578474 | 0.08928834 | 0.095899107 |
| total\_rec\_int | 1.547304418 | -0.007835539 | 0.297498476 |
| total\_rec\_late\_fee | 18.52784982 | 6.641960535 | -13.71893815 |
| last\_pymnt\_amnt | 4.070130632 | -0.21064159 | -0.111192439 |
| tot\_coll\_amt | 275.5158688 | 2.092388527 | -1.157562675 |
| tot\_cur\_bal | 2.972021144 | -0.036769912 | 0.193684065 |
| open\_acc\_6m | 1.672532644 | 0.179245539 | -0.676388802 |
| open\_act\_il | 2.995135447 | 0.003068032 | -0.140925028 |
| open\_il\_12m | 1.722697667 | 0.355067391 | -1.264250506 |
| open\_il\_24m | 1.652799851 | 0.022752852 | -0.130713446 |
| total\_bal\_il | 3.981469637 | -0.385091096 | 0.294883984 |
| open\_rv\_12m | 2.032985332 | 0.104945146 | -0.4230937 |
| open\_rv\_24m | 2.022060184 | -0.002179524 | 0.016601605 |
| max\_bal\_bc | 43.2817862 | -0.025061076 | 0.373110263 |
| total\_rev\_hi\_lim | 6.686011257 | 0.088220139 | 0.21477028 |
| inq\_fi | 2.742700892 | 0.224038053 | -0.756850482 |
| total\_cu\_tl | 3.416134864 | 0.414806164 | -0.986141176 |
| inq\_last\_12m | 2.540853318 | 0.058621741 | -0.242303871 |
| acc\_open\_past\_24mths | 1.455031497 | -0.006288317 | 0.24301794 |
| mo\_sin\_old\_rev\_tl\_op | 1.004411714 | -0.014482882 | 0.377487699 |
| mo\_sin\_rcnt\_rev\_tl\_op | 3.740485696 | 0.003614836 | -0.034645797 |
| mo\_sin\_rcnt\_tl | 5.088484604 | 0.001404619 | -0.051479056 |
| mort\_acc | 1.807777493 | 0.174081093 | -0.509915075 |
| num\_accts\_ever\_120\_pd | 5.999255651 | 1.44564769 | -4.218003803 |
| num\_actv\_bc\_tl | 1.476574326 | 0.003496646 | 0.174126296 |
| num\_actv\_rev\_tl | 1.568915012 | 0.004604149 | 0.152692588 |
| num\_bc\_sats | 1.680129016 | 0.003923152 | 0.124748862 |
| num\_bc\_tl | 1.557366484 | 0.000281125 | 0.135639105 |
| num\_il\_tl | 2.171158344 | -0.00470683 | 0.128666752 |
| num\_op\_rev\_tl | 1.555773295 | 1.49E-06 | 0.094011133 |
| num\_rev\_accts | 1.498165203 | -0.004426855 | 0.10226727 |
| num\_rev\_tl\_bal\_gt\_0 | 1.488568728 | 0.004293439 | 0.167131822 |
| num\_sats | 1.290073719 | 0.001817245 | 0.159582644 |
| num\_tl\_op\_past\_12m | 1.484488035 | -0.006500112 | 0.041920537 |
| pub\_rec\_bankruptcies | 2.569743188 | 2.328753423 | -11.82738632 |
| tot\_hi\_cred\_lim | 2.845617477 | -0.023013352 | 0.160005933 |
| total\_bal\_ex\_mort | 3.735110538 | 0.077702767 | 0.266645986 |
| total\_bc\_limit | 3.772469341 | 0.053886258 | 0.322344469 |
| total\_il\_high\_credit\_limit | 2.902536446 | -0.52458516 | 0.337406167 |
| il\_util | 0.675049519 | 250.7319275 | 8.472135812 |
| mths\_since\_recent\_inq | 0.805199351 | 5.380286868 | 8.472135812 |
| mths\_since\_rcnt\_il | 3.528819033 | 261.5932406 | 8.472135812 |
| mo\_sin\_old\_il\_acct | 0.30752551 | 101.5447327 | 8.472135812 |
| bc\_util | 0.063158632 | 238.0057717 | 8.472135812 |
| percent\_bc\_gt\_75 | 0.737423152 | 2.271636893 | 8.472135812 |
| bc\_open\_to\_buy | 3.214708802 | 287.1626323 | 8.472135812 |
| mths\_since\_recent\_bc | 3.685148938 | 260.1948526 | 8.472135812 |
| dti | 24.30497184 | 74.59678934 | 8.472135812 |
| revol\_util | 0.251091126 | 287.8559082 | 8.472135812 |
| all\_util | -0.08059644 | 371.6263114 | 8.472135812 |
| avg\_cur\_bal | 3.740417201 | 381.2538767 | 8.472135812 |
| pct\_tl\_nvr\_dlq | -2.523451004 | -0.923689079 | 8.472135812 |

**5. BIVARIATE ANALYSIS**

Now that our data has been cleaned and analysed on a univariate basis, the number of variables have reduced from 145 to 70. We've performed bivariate analysis for this finalised set of features with respect to our target variable loan\_status to see the impact the independent variable has on our dependent variable.

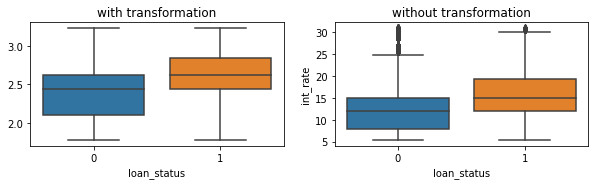
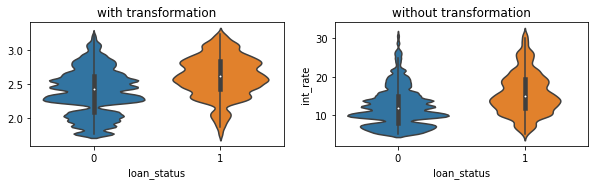
Using graphical methods:

1. As our target variable is a binary class categorical variable, so we have used the boxplot and violin plot between the continuous independent variable and target variable to understand the relationship between them.

We have used the following function:



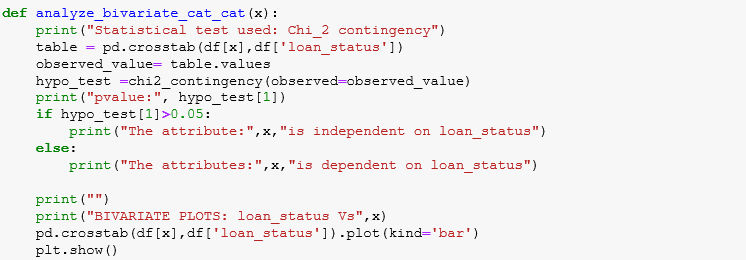
Results: Eg - int\_rate (continuous independent) Vs loan\_status (target)



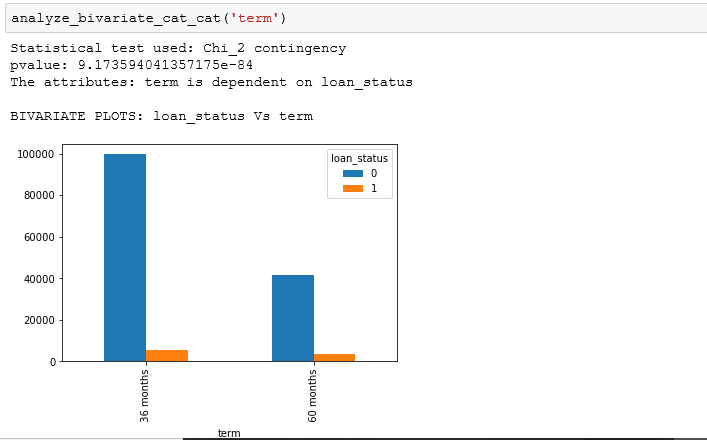
Conclusion: Our graphical analysis indicates unequal spread between interest rate for defaulters and interest rate for non-defaulters. The central tendencies of interest rate are higher for defaulters data than for non-defaulters data. Which indicates the interest rate in general is higher for defaulters than non-defaulters.

1. As our target variable is a binary class categorical variable, so we have used the barplot between the categorical variable and target variable to understand the relationship between them.

Function:



Results: Eg: term (categorical independent variable) Vs loan\_status (target variable)



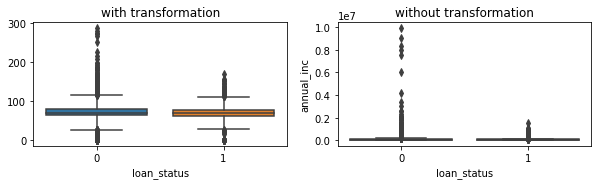
Conclusion: The plot indicates a higher population of borrowers who have not defaulted on the loan compared to the borrowers who have defaulted the loan. On a general demographic there are more borrowers who chose a term plan of 36 months over 60 months.

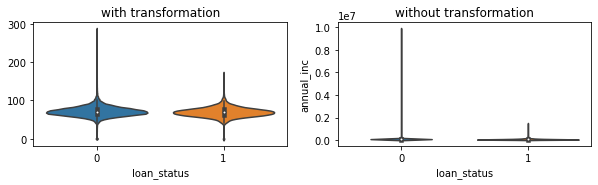
Consensus: The graphical analysis indicates more borrowers have chosen 36months term plan over 60 months. This might be due to multiple reasons but the main reason being that borrowers would prefer to be charged less interest over their term, which is automatically higher for longer terms than for shorter repayment terms. Our statistical test has given us a pvalue less than 0.05 indicating that term does have an affect on our loan\_status.

This section involves the analysis of our independent variables with our target variable loan\_status. We’re first going to consider few most important variables that impact loan approval and in turn loan risk of a borrower, which are income, DTI, interest rate and term

1. ANNUAL INCOME AND LOAN STATUS (annual\_inc, loan\_status):

Categorical vs Numerical feature analysis





Statistical Test:

KruskalResult(statistic=204.16540408157923, pvalue=2.575686846725997e-46)

MannwhitneyuResult(statistic=716033297.0, pvalue=2.575691394495958e-46)

Here we have compiled all the comparison metrics derived for our bivariate analysis between annual income and loan status including the statistical analysis of the two variables.

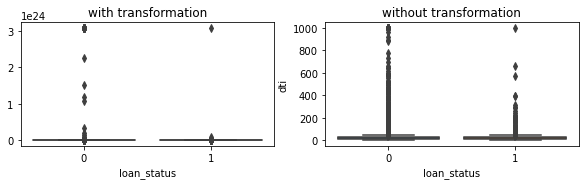
Conclusion: From the plots generated it indicates that the spread of annual income is similar irrespective of loan status being default or non-default. But the range of annual income is definitely wider and higher for non-defaulters even if the central tendencies are pretty stationary for distribution of income with respect to loan status. It’s clearly shown here that few income levels are not indicated in loan defaulters data.

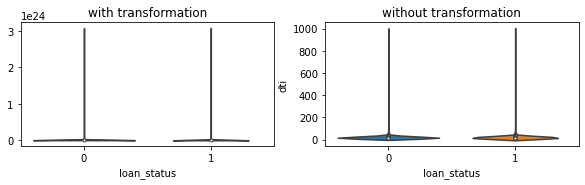
Our correlation matrix shows negative correlation between loan status and income which indicates that the chance of a person being a loan defaulter grows with decrease in annual income.

Consensus: The results indicate the exact business standpoint we’ve discussed before. That our annual income affects a credit risk. The higher the annual income the lower the chances of getting into any delinquencies or defaulting for the borrower. Our statistical tests generated pvalues less than 0.05 indicating that annual income is significant to prediction of loan status.

1. DTI AND LOAN\_STATUS:

Categorical vs Numerical analysis





Statistical Test:

KruskalResult(statistic=nan, pvalue=nan)

MannwhitneyuResult(statistic=646616787.0, pvalue=0.005368382574994331)

Here we have compiled all the comparison metrics derived for our bivariate analysis between dti and loan status including the statistical analysis of the two variables.

Conclusion:

The plots indicate a similar distribution for both defaulter and non-defaulter borrower DTI data. Even the range doesn’t differ. Indicating that DTI has no affect on loan\_status. Or atleast that there’s significant impact that DTI delivers on our loan status. Whatever influence DTI has on loan status might be minimal. Our correlation value of 0.006 indicates the same

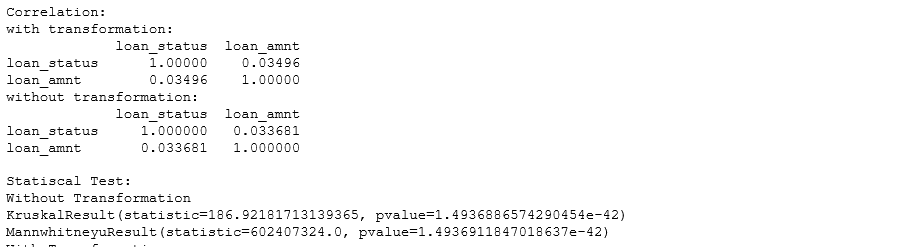
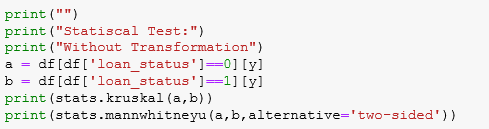
Consensus:

Our graphical output and correlation metrics deviate from our business understanding. Which concluded that DTI is a highly defining variable for credit risk. This might be because of few reasons including that there’s not a good spread of DTI data in our dataset. Our statistical tests indicates a pvalue of 0.005 which is borderline indication of DTI’s significance to our target variable. While it does indicate that DTI and loan status are related, the pvalue is not low enough for this to be considered a strong relationship.

Similar procedure was followed for all the continuous and categorical variables

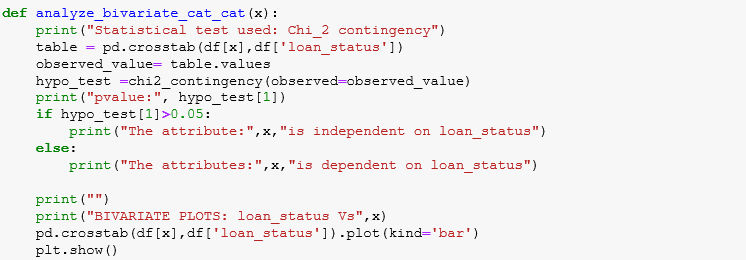
1. UNI FEATURE SELECTION USING STATISTICAL TESTS

From the above bi-variate analysis , it has been inferred that all the columns are non-parametric and unpaired data. Thus we used the Manwhitneyu (t-test) and kruskal ( ANOVA) between the continous variable and target variable.

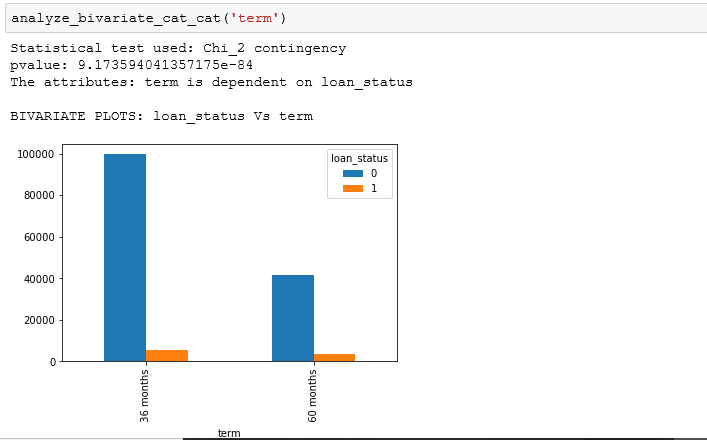


Similarly, we have written a function to find the relationship between the independent categorical variable and target variable using the statistical test chi-square independence of attributes.

Function:



Results : Eg: term (categorical independent variable) Vs loan\_status (target variable)



|  |  |
| --- | --- |
| total\_acc | Stronger relationship |
| total\_rec\_int | Stronger relationship |
| last\_pymnt\_amnt | Stronger relationship |
| tot\_cur\_bal | Stronger relationship |
| total\_bal\_il | Stronger relationship |
| max\_bal\_bc | Stronger relationship |
| total\_rev\_hi\_lim | Stronger relationship |
| acc\_open\_past\_24mths | Stronger relationship |
| mo\_sin\_old\_rev\_tl\_op | Stronger relationship |
| mo\_sin\_rcnt\_rev\_tl\_op | Stronger relationship |
| mo\_sin\_rcnt\_tl | Stronger relationship |
| num\_sats | Stronger relationship |
| num\_tl\_op\_past\_12m | Stronger relationship |
| tot\_hi\_cred\_lim | Stronger relationship |
| total\_bal\_ex\_mort | Stronger relationship |
| total\_bc\_limit | Stronger relationship |

In this way we found the independent variables which have stronger relationship with the target variable usng the pvalues obtained from the statistical results. Lower the p-value , stronger the relationship. Yet this is not a robust feature selection method .

**7.BASELINE MODEL BUILDING**

As our problem statement is binary classification problem indicating that target variable is categorical in nature so we used the following 9 classification algorithms. For the base model building, we have considered all the 72 independent features that have been finalized after our exploratory data analysis and outlier treatment.

|  |  |
| --- | --- |
| Encoding | One Hot Encoding |
| Imputation Method : | Mean, Mode. Median |
| Transformation | yeo-jhonson |
| Scaling | Standard Scaler |
| Features | All Features |
| Scoring | "roc\_auc" |

For all the below models we have used the above metrics and found out the following results.

Metric used for deciding the model : Recall

Why Recall ?

False Negatives are the values which the model predicted negative class(0), that means model predicted the person will not default , but this prediction is false meaning in reality he defaulted , so we have to reduce False Negatives because these people will be a contributing for loss of the company. Recall is defined as TP / (TP + FN) which indicates that while we reduce FN our Recall should increase, thus Recall should be the metric for our model.

Why not Precision?

Precision is defined as TP/(TP+FP) .False Positives are the values which the model predicted positive class(1) , that means model predicted the person will default but prediction is false,meaning in reality he did not default.

But if our problem statement was something like we had to give offers or special benefits for people who did not default , then we could lose these customers , then in that case precision would be our parameter for our model

90 10 % ,

**random**