**Loan Default Prediction – Milestone Report**

1. **Problem Statement**

**Why useful Questions?**

**For Whom?**

**Statement: *Predicting which loans that are most likely to default***

Predicting which loans are most likely to default

1. **Dataset description**
2. **EDA**

Given that this dataset contains 145 features, it would be extremely time consuming analyzing each feature. This section was structured by analyzing significant features while asking a series of questions and answering them with visuals. Details will be provided regarding the following:

* Loan amounts
* Terms
* Interest rates
* Installments
* Grades and subgrades
* Employment Analysis
* Home Ownership
* Annual income
* Loan Status
* Purpose of Loan
* Location Analysis of Loans
* How have features changed over time?
* Are there features with significant correlations?
  1. ***Loans Amount***

1. **What does the distribution of Loans look like?**

**A screenshot of a cell phone

Description automatically generated**

* The distribution is skewed to the right
* The most common loan is $10,000

By plotting the ECDF we can see where most of the distribution lies

1. **What does ECDF of Loan Values look like?**

A close up of a map

Description automatically generated

* 50% of Loans are less than USD 13,000
* 90% of loans are less than USD 29,000
* The largest loan is at USD 40,000 with the lowest loan at USD 500.
* There is a spike in loans at round numbers 10,000 ,20,000, 25,000 etc

By plotting the ECDF we can see where most of the distribution lies

* 1. ***Term of the Loans***

1. **How many loans are 36 months and how many are 60 months?**

* 71% of issued loans are 36 months
* 29% of issued loans are 36 months

1. A screenshot of a social media post

   Description automatically generated**How do Loan amounts differ between the 2 terms?**

The following observations were seen:

* Mean for 60 months: $20,000
* Mean for 36 months: $10,000
* Both terms have an equal range of loan values
* 36 month IQR: $6,000 AND $17,000
* 60 month IQR:$15,000 AND $25,000
* 36 month loans has a few outliers outside the 1.5 IQR whisker.
  1. ***Interest Rates***

1. A picture containing computer

   Description automatically generated**What does the distribution of interest rates look like?**

* The most common interest occurs at around 12-13%
* The interest rate distribution tails to the right, like loan amounts.

1. **What does ECDF of interest rates look like?**

A close up of a person

Description automatically generated

* 50% of Loans are issued at ≤13%
* 90% of Loans are issued at ≤20%
* The maximum interest rate is 33%
* The minimum interest rate occurs at 5%

1. **How does the term of the loan affect interest rate?**

A picture containing screenshot

Description automatically generated

* **36 months**
  + The IQR is between 7 and 14%.
  + The whisker edges are between 5% and 23%
  + The range is between 5% and 33%
* **60 months**
  + The IQR is between 13% and 18%.
  + The whisker edges are between 5% and 27%
  + The range is between 5% and 33%

In order to see a more complete picture of the loans, the loan amount will be integrated in the next question.

1. A screenshot of a cell phone

   Description automatically generated**How are loan amounts, interest rate and term change with respect to each other?**

* **60 months**
  + $9,000 seems to be where loans start getting issued. Mostly at 15% Interest rate.
  + Interest rates do not exceed 27%.
  + Interest rates range from 7% to 27%. The higher the loan value, the lower the interest rates range.
  + The highest concentration of loans are between 10,000 & 20,000
  + The highest concentration of interest rates is between 10% and 17%.
  + The mean loan is ~$20,740 with a std of $8090 and a mean IR of 15.9%.
* **36 months**
  + The highest concentration of loans are between $8,000 & $12,000
  + The highest concentration of interest rates is between 5% and 12%
  + Loans start to significantly decrease after 20,000.
  + The mean loan is ~$12,750 with a std of $8560 and a mean IR of 12%.

In general, the higher the loan the lower the interest rate. A $10,000 loan over 36 months tends to have a lower interest rate than a $10,000 loan over 60 months.

* 1. ***Installments***
     1. **How are Installments and Loan values related?**

***A screenshot of a computer

Description automatically generated***

**Observations:**

* There is a linearly increasing ceiling for installments which indicates a direct relationship between installments and loan amounts.
* As the loan increase the variability of the installment value increases but does it increase by the same proportion?

Example:

* The maximum installment issued for a $10,000 loan is around $500 and a minimum installment of $100. A $400 range.
* The maximum installment issued for a $40,000 loan is around $1,750 and a minimum installment of $750. A $1,000 range.
* The loan value increased by 4x but the range only increased by 2.5x.
  + 1. **A screenshot of a video game

       Description automatically generatedHow do Installments differ by Term?**
* 36 months has a range of installments between 0 and $1750
* 60 months has a range of installments between 0 and $1300
* 36 months has an IQR between $230 and $550.
* 60 months has an IQR between $300 and $650.

36 months has a wider range of installments than 60 months but the on average 60 months pay more every installment. Let's integrate loan amount in the above graph.

In order to see a more complete picture of the loans, installment and term will be integrated in the next question.

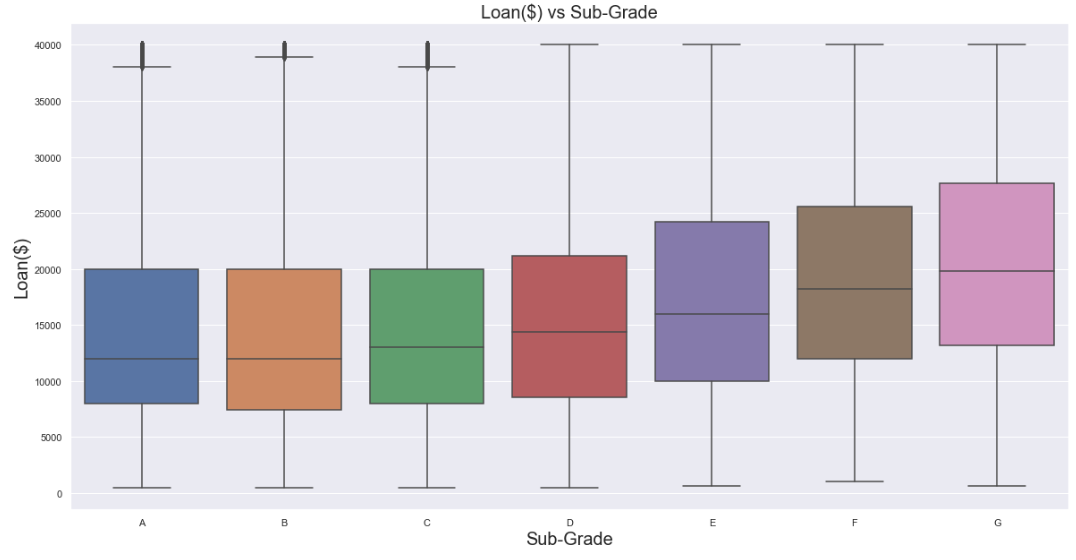
1. **How do Installments differ by Term and loan amount?**

A screenshot of a map

Description automatically generated

For the same loan amount, 36 months have higher installments than 60 months as can be seen above.

* 1. ***Grade and Subgrade***
     1. ***A screenshot of a video game

        Description automatically generated*How does interest rate change with grade of the loans?**
* Interest rate is highly dependant on the grade assigned.
* There exists a few outliers for each grade at very low interest rates.
  + - ****How do loans change with grade of the loans?**

Grade and loan amount start having a correlation starting from grade D.

1. ***A screenshot of a cell phone

   Description automatically generated*How does interest rate change with subgrade?**

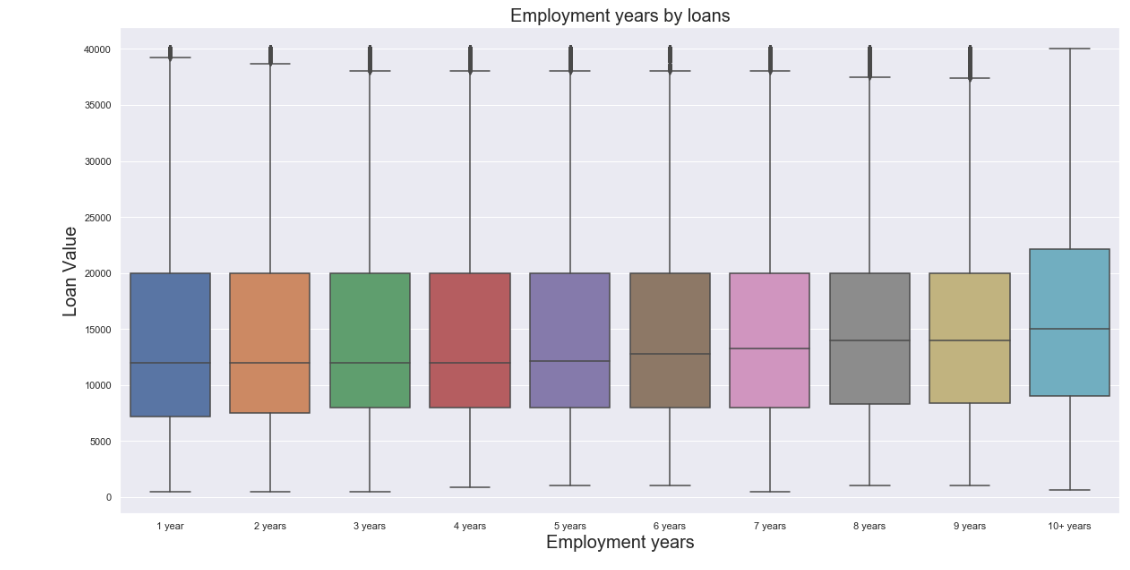
As subgrade deteriorates, not only does the interest rate increase, but the range increases as well. Example: A B4 loan has a range between 10-13%. An E4 loan has a range between 14-26%. This shows a huge increase in range and variability.

* 1. ***Employment Information***
     1. **How long have the loan holders been in employment?**

A picture containing drawing

Description automatically generated

* 35% of Loanees have been working for 10+ years
* 65% have been working less than 10 years as seen in the above graph.
  + 1. **Do loan values change based on employment years?**



* The IQR for those employed for 10+ years is between $9,000 and $23,000 and there are not any significant outliers as all loans fall below 1.5\*IQR.
* The IQR for those working for less than 10 years is between USD8,000 and USD 20,000.
  1. ***Home Ownership***
     + 1. **What kind of home ownership do clients have?**

***A screenshot of a cell phone

Description automatically generated***

* 50% of clients have a mortgage
* 40% of clients rent a property.
* 9% already own a property.
* 1% are unidentified.
  + - 1. A close up of text on a white background

         Description automatically generated**How does loan amount change with homeownership?**
* Mortgage holders tend to hold higher loans and have the widest variability with a mean of $15,000.
* Owners seem to have a similar mean to renters.
* Owners have higher variability than renters/
  1. ***Loan Status Analysis***
  2. ***A screenshot of a cell phone

     Description automatically generated*How are loan statuses distributed?**
  + 80-85% of loans are in good shape (Fully Paid or Current)
  + 10-12% of loans are charged off
  + Cannot make an accurate estimation on prediction of bad loans.
  1. **How are bad loan statuses distributed?**

Bad Loans were calculated to be 1.53% of total loans and are classified as the following:

* + - 1. ***In Grace Period***
      2. ***Late (16-30 days)***
      3. ***Late (31-120 days)***
      4. ***Default (120+ days)***

***A screenshot of a cell phone

Description automatically generated***

* 63-65% of late loans are between 31-120 days late.
* 20-25% of late loans are in grace period (1-16 days)
* 10-12% of late loans are 16-30days late.
* Default loans are negligible.
  1. **A screenshot of a cell phone

     Description automatically generatedHas term distribution changed between all loans and late loans?**
* 60 months loans made up 28% of all loans but made up 42% of late loans.
* 36 months loans made up 72% of all loans but made up 57% of late loans.
* The term of the loan does indeed affect the lateness of the loan.

The following was observed from the given data:

|  |  |  |
| --- | --- | --- |
|  | ***Average Loan Amount*** | ***Average Interest Rate(%)*** |
| ***All Loans*** | ***$15,000*** | ***13.1%*** |
| ***Late Loans*** | ***$17,000*** | ***15.6%*** |

* 1. ***Loan Purpose Analysis***

1. **What purposes are used for the loans?***A screenshot of a cell phone

   Description automatically generated*

* 55-60% of loans are for debt consolidation.
* 17-23% of loans are used to pay Credit card
* 7-8% of loans are used for home improvement purposes
* There is a decrease in loans for Credit card from 23% for all loans to 17% late loans.

1. ***A picture containing drawing

   Description automatically generated*How does the purpose change the loan amount?**

* Loan amounts is highest for small businesses.
* Loan amounts is lowest for vacations.
  1. ***Location Analysis***
     + 1. **How do number of loans vary by state?**

***A screenshot of a cell phone

Description automatically generated***

* California has the most amount of issued loans(13%).
* Iowa has the lowest amount of issued loans at close to 0%
  + - 1. **How does the mean loan value vary by state?**

**A picture containing screenshot

Description automatically generated**

* California has the 8th highest mean loan value.
* Alaska has the highest mean loan

In order to combine the effect of loan value and number of loans, what does the sum of all loans for each state look like?

1. **How does the sum of all loans vary by state?**

* Like the number of loans with CA state as the #1.
* The order is mildly different.

A picture containing screenshot

Description automatically generated

* 1. ***How have features changed with time***
     + 1. **A picture containing photo, kitchen, white, table

          Description automatically generatedHow does Loan count change across the years?**

**Observations**

* It took 6-7 years(2007-2013) for the number of loans to go from 0 to 10,000. The increase was steady with no fluctuations month to month.
* It took 2-3 years (2013-2015) for the number of loans to go from 10,000 to 20,000.The increase was steady with no fluctuations month to month.
* It took 1-2 year(2015-2016) year to go from 20,000 to 60,000 loans. But there was severe fluctuations month to month.
* There was a peak of loans in 2016 but then decreased rapidly to 30,000 issued loans.
  + - 1. **A close up of a map

         Description automatically generatedHow does the mean loan change across the years?**

The mean loan value increased from 2007 to 2012, flattening at around $15-16,000 per loan.

* + - 1. **A close up of a map

         Description automatically generatedHow does the interest rate change across the years?**
* Interest rates were steady between 2008 and 2012 ranging between 11-13%.
* Interest rates increased from 2012 to 2014 reaching a peak of 15%.
* They then decreased to 12% reaching a low in 2016.

1. **How does Loan count change with interest rate?**

**A screenshot of a computer

Description automatically generated**

There are two observations evident.

* 1st Trend: We see negative relationship from 12% to 15% for most of the months.
* 2nd Trend: We see a flat relationship between 9-13% for a very low number of loans.

1. **How does mean interest rate change with Loan Value?**

**A screenshot of a computer

Description automatically generated**

There is a strong positive relationship between loan value and interest rate for loans between 4,000 and 16,000.

1. ***A screenshot of text

   Description automatically generated*How has the term length change across the years?**

The pattern of change for both terms of loans are similar but there is an evident difference of distribution as examined previously.

* 1. ***Are there significant correlations between the numeric variables in the data frame?***

A screenshot of a social media post

Description automatically generatedThe aim of this question is to visualize if there are significant correlations with pearson correlation coefficients with magnitudes greater than 0.7.

There does exist multiple features with strong correlations that could be helpful in predicting whether a loan will be bad.