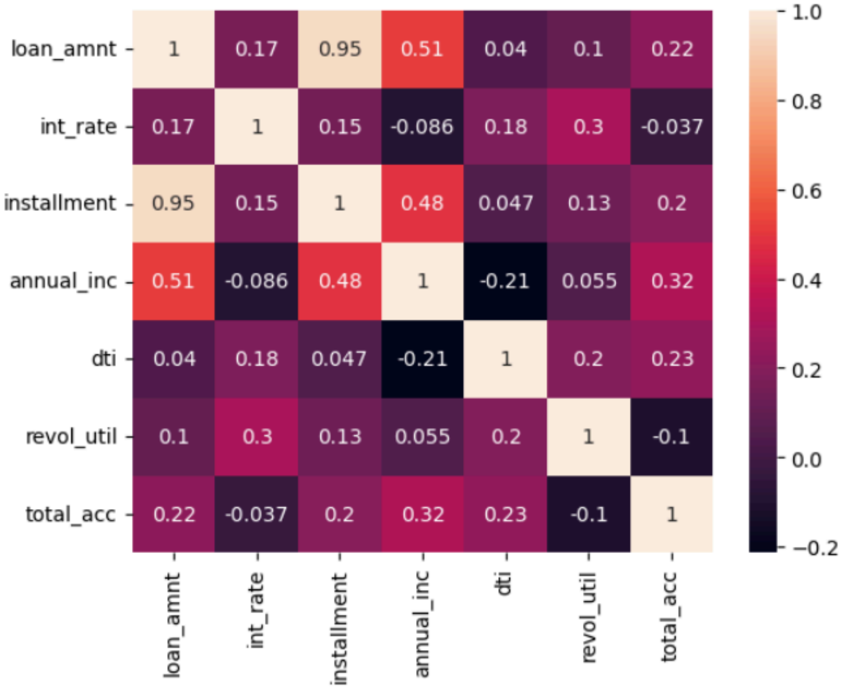
Questionnaire

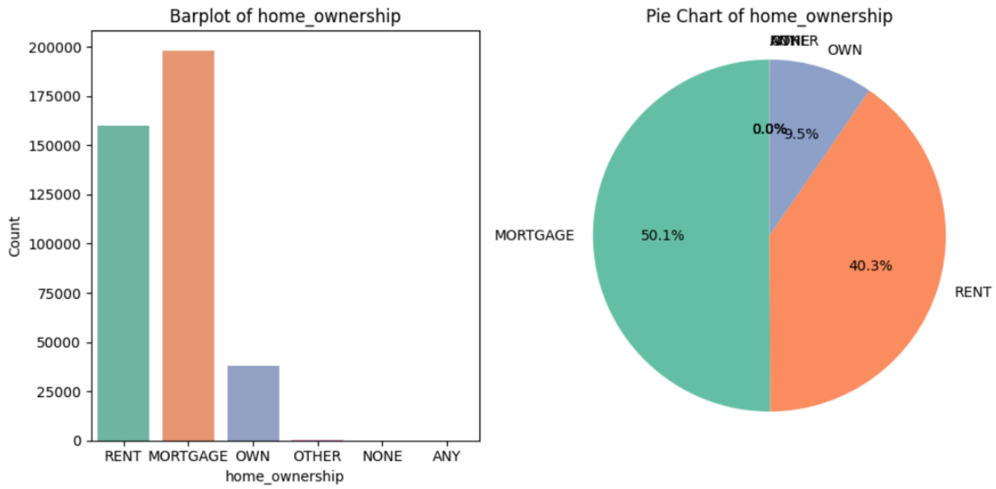
1. What percentage of customers have fully paid their Loan Amount?
   80.4% customers have fully paid their loans, 19.6% customers have been charged off (not paid). The purpose of the case study would be to predict, if Customer would be able to pay back their loan. The dataset is imbalanced

2. Comment about the correlation between Loan Amount and Installment features.

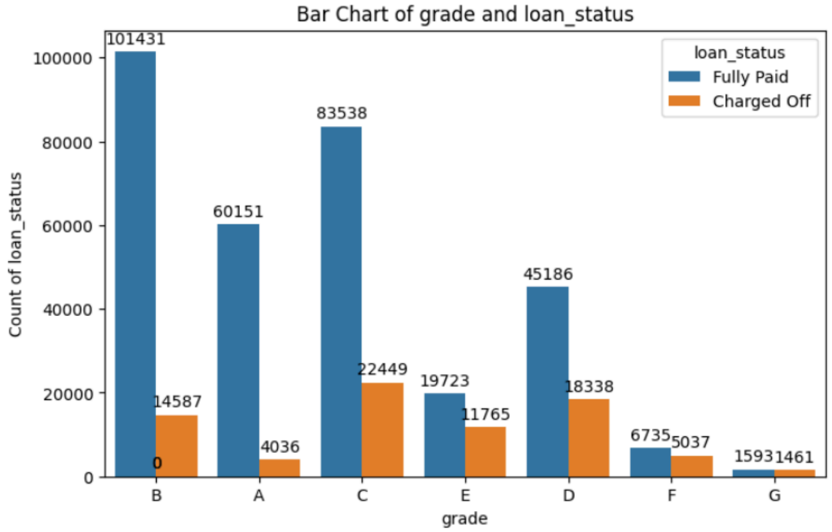


   There is a high positive (0.95) correlation between loan amount and installment. This suggests a higher amount of loan would certainly translate to a higher amount of installments (monthly payment owed).

3. The majority of people have home ownership as _______.
   50% customers own a mortgage home, followed by rent followed by Own. Customers on rent, 36212÷(36212+123578) is approximately ~22% Customers have not paid their loans.



4. People with grades ‘A’ are more likely to fully pay their loan. (T/F)



Customers with Grade A are likely to pay back loans,. 60151 customers have paid back loans, only 4036 have been charged off. This is the lowest charged off count among prominent grades.

5. Name the top 2 afforded job titles.

emp_title has 173105 unique values, usage of this feature would likely complicate the model, so was not used.

6. Thinking from a bank's perspective, which metric should our primary focus be on..
   1. ROC AUC

2. Precision
3. Recall
4. F1 Score

Precision will be the primary focus. As precision = TP / (TP + FP). We want to reduce FP to minimum, as FP would represent Customers, that model classified are potential candidates, but are actually not worthy of paying back Loans. Such Customers would turn losses to the bank.

7. How does the gap in precision and recall affect the bank?
   precision = TP / (TP + FP), recall = TP / (TP + FN).
   It does not affect Bank much for rejecting slightly doubtful cases (possible False Negatives), but granting Loan to incapable Customers will result in a huge loss, ie False Positives. So since for stricter scrutiny precision is preferred over recall.

8. Which were the features that heavily affected the outcome?
   Pin Code is a distinctly dominant feature, all coefficients are at least ~20 times smaller than Pin Code.

```
{'pin_code': 9.12913409779421,
 'grade': 0.4302566939515996,
 'annual_inc': 0.2349782709173531,
 'issue_year': 0.23300821083154827,
 'open_acc_54-90': 0.068322197154072,
 'purpose': 0.056952788683906214,
 'revol_bal_> 19620': 0.05243962565635118,
 'initial_list_status': 0.052166846851082034,
 'state': 0.043793841637418916,
 'pub_rec_bankruptcies_1': 0.043676495646885465,
 'revol_bal_11181-19620': 0.038217602657513415,
 'total_acc': 0.03640133908946984,
 'mort_acc_3-10': 0.030865611906146992,
 'issue_month': 0.02573921687401446,
 'revol_bal_6025-11181': 0.02439680542262263,
 'earliest_cr_year': 0.015685169118199525,
 'earliest_cr_month': 0.01320370929982393,
 'emp_length': 0.009572714526267087,
 'mort_acc_10-20': 0.008859209701954907,
 'mort_acc_20-34': 0.0007863697218385342,
 'home_ownership_OTHER': -0.0007277366349155 02,
 'pub_rec_bankruptcies_More_than_1': -0.005349402884082261,
 'open_acc_42-54': -0.005769910191269469 5,
 'home_ownership_NONE': -0.009592482223329235,
'pub_rec_[10, 30)': -0.010524053433499317,
'pub_rec_[5, 10)': -0.0155586728816656,
'pub_rec_[30, inf)': -0.016746687881941873,
'open_acc_30-42': -0.01876183192782853б,
'home_ownership_OWN': -0.03732878485645253,
'open_acc_18-30': -0.04338162289328818,
'verification_status_Verified': -0.04858446046846344,
'pub_rec_[1, 5)': -0.05279235369356111,
'verification_status_Source Verified': -0.05569029901014546,
'revol_util': -0.10883900821837027,
'installment': -0.11729597527950333,
'home_ownership_RENT': -0.11890524265288026,
'dti': -0.1704912363519588,
'term': -0.2189137522492677}
```

9. Will the results be affected by geographical location? (Yes/No)
   Yes, Pin Code Extremely dominates all other (non-geographical) features, please see previous answer for details.

Overall review and Recommendations

Distribution & Outlier Evaluation

1. loan_amt

- All outliers are almost on upper whisker
- Kurtosis of almost 0 signify no outliers
- Therefore, outliers can be assumed to be not present.
- The distribution is moderately right skewed with 0.7 as skewness, suggesting customers prefer smaller loans.

2. int_rate

- Interest rates are usually in range of 10.5 to 16.5%
- Kurtosis is -0.14%, signifying absence of outliers.
- All outliers are nearby upper whisker and total outliers are 0.95%
- Overall, outliers can be assumed to be not present.
- The distribution is about symmetrical with a skewness of 0.42, might be influenced by outliers.

3. Installment

- Skewness of 0.98 confirms right skewness, so most of the installments are on the lower side.
- Kurtosis of 0.78, signifies absence of large outliers in distribution.
- All outliers are concentrated in a range above upper whisker and total outliers are 2.84%
- Overall, outliers can be assumed to be not present.
- Max installment is 1533.81, so installments are of lower values.

4. annual_inc

- Percent Outliers in annual_inc is 4.22%
- Annual incomes exist in the range of 0.0 to 8706582.0. It is a huge range, and outliers are distributed sparsely at extreme points. Also, Kurtosis of 4238.55 confirms a wide range.
- Incomes are heavily right skewed, skewness is 41.04
- This feature requires binning.

5. dti

- Q3 is 22.98, and max value is 9999.0. It is an extreme range.
- There are only two points above 100. These 2 points are definite outliers.
- Only 0.07% records are outliers (275 rows). As total rows are 396030, we can skim this feature.
- Skewness of 431.05 and Kurtosis of 237923.67, are extremely large, after rechecking after outlier, statistics will be re-evaluated and decision on feature binning will be confirmed.
- Since only 275 records were outliers, they were capped at whisker points.

6. open_acc

- Open Account has Q3: 14.0 (IQR: 6.0) and max value of 90.0. There is an extreme range in this feature.
- Percent Outliers in open_acc is 2.60%. This is a significant portion.
- Skewness: 1.21 signifies high right skewness.
- Overall, this feature should be binned

7. Pub_rec

- Q1: 0.0, Q3: 0.0 and IQR: 0.0
- Min: 0.0, Max: 86.0, Mean: 0.17, Median: 0.0.
- Percent Outliers in pub_rec: 14.58%
- This is a clear case for feature binning, as this feature represents exceptional cases - negative public records

8 revol_bal

- Q1: 6025.0, Q3: 19620.0, Max: 1743266.0 - distribution is in a large range.
- Percent Outliers in revol_bal: 5.37%
- This feature will require feature binning or alternate transformation

9. revol_util

- Total Outliers in revol_util: 12, Percent Outliers in revol_util: ~0.00%
- There is 1 extreme outlier that will require removal, otherwise all data points are below 200
- All other outlier points are near the upper whisker
- Skewness: -0.07188910791723081, Kurtosis: 2.716256667706004 suggest that distribution is not skewed and there are not a lot of outliers.

10. total_acc

- total_acc is in a wide range, Q1: 17.0, Q3: 32.0, Max: 151.0.
- All outlier points are concentrated uniformly around the whisker.
- Skewness: 0.86,suggests right skewness.
- Total Outliers in total_acc: 8499, Percent Outliers in total_acc: 2.15%.
- Kurtosis: 1.20 suggests that the feature is more uniformly distributed across a wide range.
- Overall, we can assume that outliers are not present, few extreme points
- We may consider binning this feature as range

11. mort_acc

- Total Outliers in mort_acc: 6843, Percent Outliers in mort_acc: 1.73%
- Q1: 0.0, Q3: 3.0, Max: 34.0 suggest outliers
- Kurtosis: 5.30, suggests and percent outliers are prominent in this feature.
- Skewness: 1.7558645341907244
- We will perform feature binning for this feature

12 pub_rec_bankruptcies

- `Q1: 0.0, Q3: 0.0, IQR: 0.0 suggests this feature covers exceptional cases`
- `Min: 0.0, Max: 8.0, Median: 0.0, feature has a short range`
- `This feature must be feature binned`

`Total Cols - 12`

Features to bin (must)

1. Pub_rec_bankruptcies
2. Open_acc
3. Mort_acc
4. Pub_rec
5. revol_bal

Features to bin (good to have)

1. Annunal_inc (it is continuous feature, not recommended)
2. Dti

Outlier removal

1. Dti
2. Annual_income (sp case)
3. int_rate
4. Loan_amt
5. Installment (sp case)

Categorical Features

Loan_status

- 80.4% customers have fully paid their loans, 19.6% customers have been charged off (not paid).
- The purpose of the case study would be to predict, if Customer would be able to pay back their loan.
- The dataset is imbalanced

Term

- Most customers prefer 36 months of loan.
- 76.3% loans are 36 months, 23.7% loans are 60 months.
- 30033 /(30033 + 63992) = 32% of customers with a 60 month term, didn't pay their loans.

Emp_length

- 36.4% customers have been employed for 10+ years, followed by 2 years and less than 1 year (<10% each).
- This is an interesting combination of very experienced employees and new employees

Home Ownership (requires cleaning)

- 50% customers own a mortgage home, followed by rent followed by Own
- Customers on rent, 36212÷(36212+123578) is approximately ~22% Customers have not paid their loans.

Verification_status

- All verification statuses are equally distributed.
- Larger portion of customers that were not verified have paid their loans.

Purpose

- 59.2% customers take loan for Debt_consolidation, followed by credit card (21%)
- Above constitute 80% of customers

List_status

- 60% of the loans are F, other 40% and W

Application_type (requires investigation)

- 99.8% loans are Individual loans
- Only 771 loans are not individuals

Pin_code

- Customers from all pin codes contribute to 10% to 15% loans, except 11650, 33700 and 86630.
- All loans in 11650, 86630, 93700 have been charged off.
- 00813, 29597, have fully paid all loans.

Grade

- Grade B and C constitute of 55% of the loans
- Grade A and D constitutes ~16% loans
- Grade E, F, G contribute to ~10% of loans

Issue_month

- All months get similar proportion of loans, July and October have highest loans.

Issue_year

- 2013, 2014, 2015 has contributed to 75% of loans, almost equally.

Earliest_cr_year

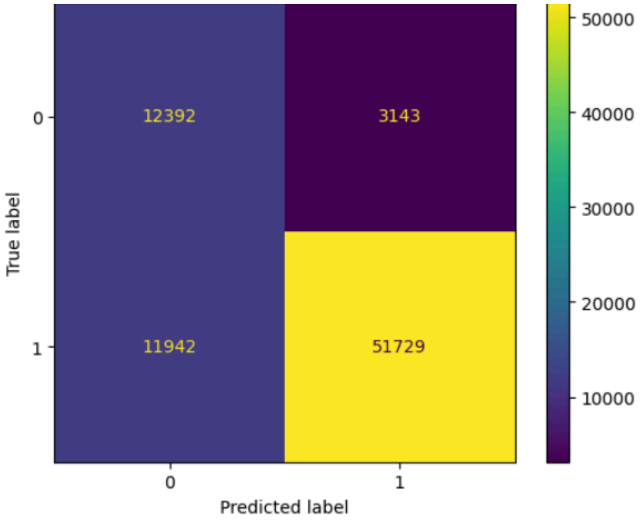
- This feature has a wide range of years

Earliest_cr_month

- Each month is contributing to 7% to 10%.

Among imbalance treatments, adjusting class weights was identified to be best. It is simpler approach providing best results. With SMOTE we observed that train set was slightly underperforming and test was better. This is due to imbalance treatment done on train set, so test set became easier for the model.

94% precision obtained on Test set. Results on train set were slightly lower, as smote was applied on train set, so Test set is performing better.

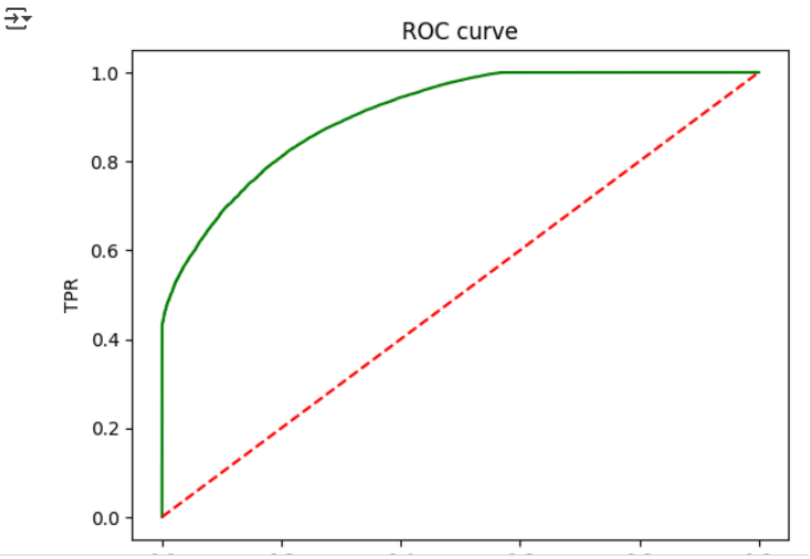


```
[282] prec = precision_calc(conf_matrix)
      print(f"Precision of Final Model: {prec}")
```

```
Precision of Final Model: 0.9427212421635807
```

Model is performing with 94.3% accuracy.

We achieved a good ROC



Overall coefficients, suggest Pin Code is highly significant.

{'pin_code': 9.12913409779421,

'grade': 0.4302566939515996,

'annual_inc': 0.2349782709173531,

'issue_year': 0.23300821083154827,

```
'open_acc_54-90': 0.068322197154072,

'purpose': 0.056952788683906214,

'revol_bal_> 19620': 0.05243962565635118,

'initial_list_status': 0.052166846851082034,

'state': 0.043793841637418916,

'pub_rec_bankruptcies_1': 0.043676495646885465,

'revol_bal_11181-19620': 0.038217602657513415,

'total_acc': 0.03640133908946984,

'mort_acc_3-10': 0.030865611906146992,

'issue_month': 0.02573921687401446,

'revol_bal_6025-11181': 0.02439680542262263,

'earliest_cr_year': 0.015685169118199525,

'earliest_cr_month': 0.01320370929982393,

'emp_length': 0.009572714526267087,

'mort_acc_10-20': 0.008859209701954907,

'mort_acc_20-34': 0.0007863697218385342,

'home_ownership_OTHER': -0.000727736634915502,

'pub_rec_bankruptcies_More_than_1': -0.005349402884082261,

'open_acc_42-54': -0.0057699101912694695,

'home_ownership_NONE': -0.009592482223329235,

'pub_rec_[10, 30)': -0.010524053433499317,

'pub_rec_[5, 10)': -0.0155586728816656,

'pub_rec_[30, inf)': -0.016746687881941873,

'open_acc_30-42': -0.018761831927828536,

'home_ownership_OWN': -0.03732878485645253,

'open_acc_18-30': -0.04338162289328818,

'verification_status_Verified': -0.04858446046846344,

'pub_rec_[1, 5)': -0.05279235369356111,

'verification_status_Source Verified': -0.05569029901014546,

'revol_util': -0.10883900821837027,

'installment': -0.11729597527950333,

'home_ownership_RENT': -0.11890524265288026,
'dti': -0.17049123635195588,
'term': -0.2189137522492677}
```