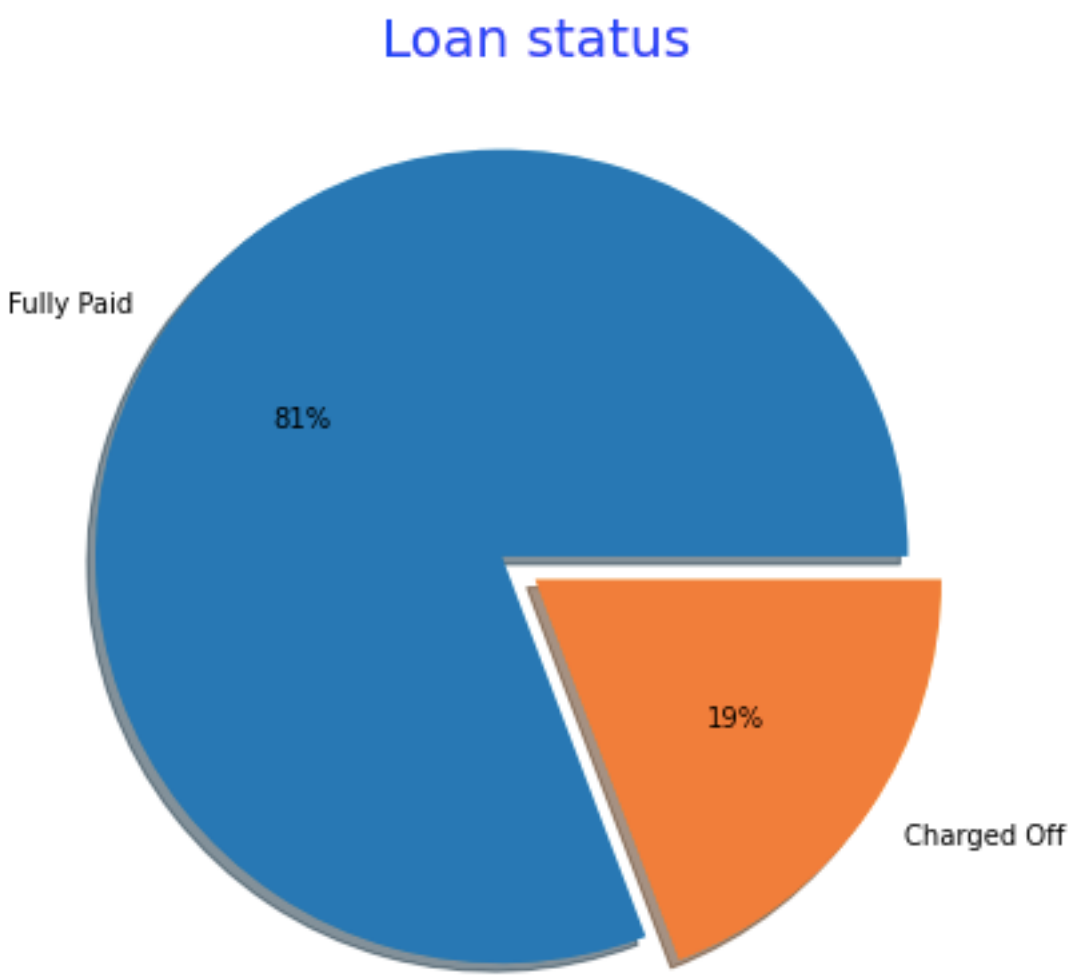
LoanTap Logistic Regression Insights and Recomendations

INSIGHTS:

- **Data:** LoanTap dataset includes over 3.96 million data points with 27 features, exploring loan characteristics and borrower details.
- **Loan Distribution:** Loan amounts range from 500 to 4000, with most borrowers having employment exceeding 10 years and opting for shorter 30-month terms.
- **Target Variable:** The target variable is "loan_status," indicating whether the loan was fully paid or not (charged off).
- **Imbalanced Classes:** The dataset exhibits class imbalance, with 80% of loans being fully paid and 20% charged off.
- **Feature Relationships:** Strong positive correlation between loan amount and installment amount is expected.
- **Model Performance:** The model achieves 81% accuracy on both training and test data, but the F1 score (0.12) and Precision-Recall AUC (0.35) indicate room for improvement, especially in identifying loan defaults (class 1).
- **Potential Improvements:** Hyperparameter tuning and exploring alternative models could potentially enhance the model's ability to predict loan defaults.

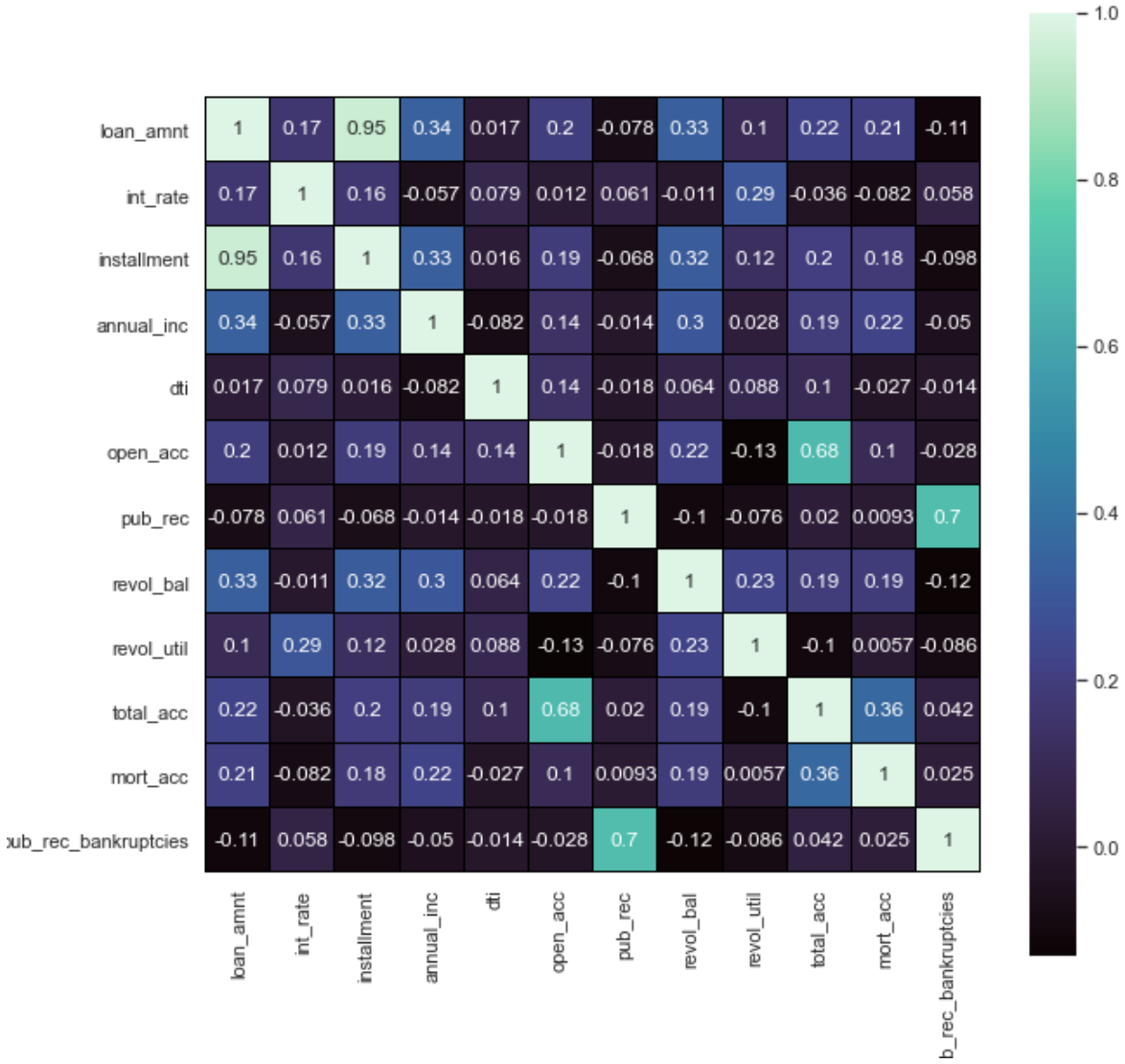
What percentage of customers have fully paid their Loan Amount?

From the below plot it is known that 81% of the customers had fully paid the loan and 19% didn't.



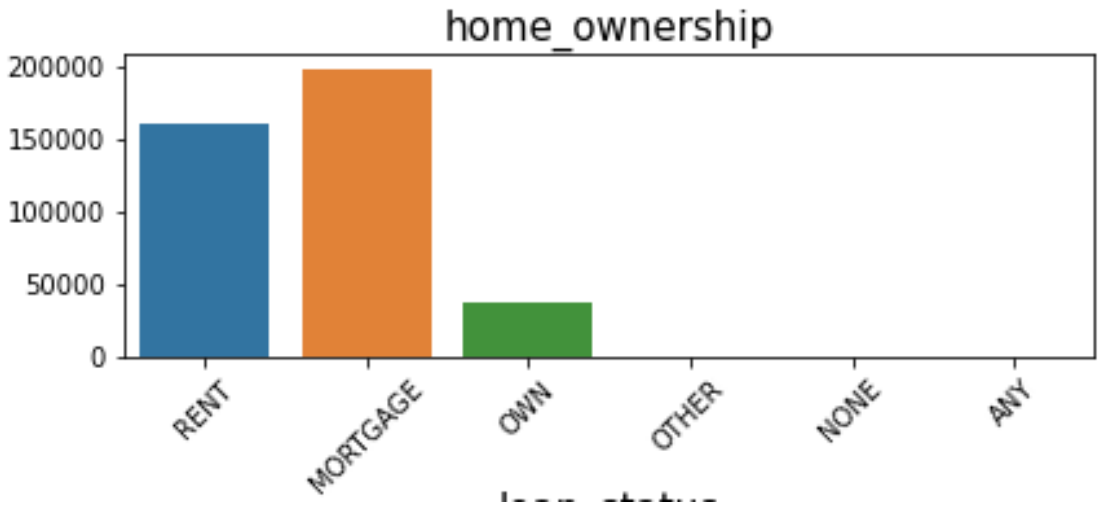
Comment about the correlation between Loan Amount and Installment features.

From the below plot we can say that feature 'installment' and 'loan_amnt' are direclty proportional and more correlated. If loan amount increases the installment amount also increases and if loan amount decreases the installment decreases. The heatmap says that 'installment' and 'loan_amnt' are 95% correlated.
Where as 'inst_rate' and 'loan_amnt' are not correlated. their correlation value is 17%



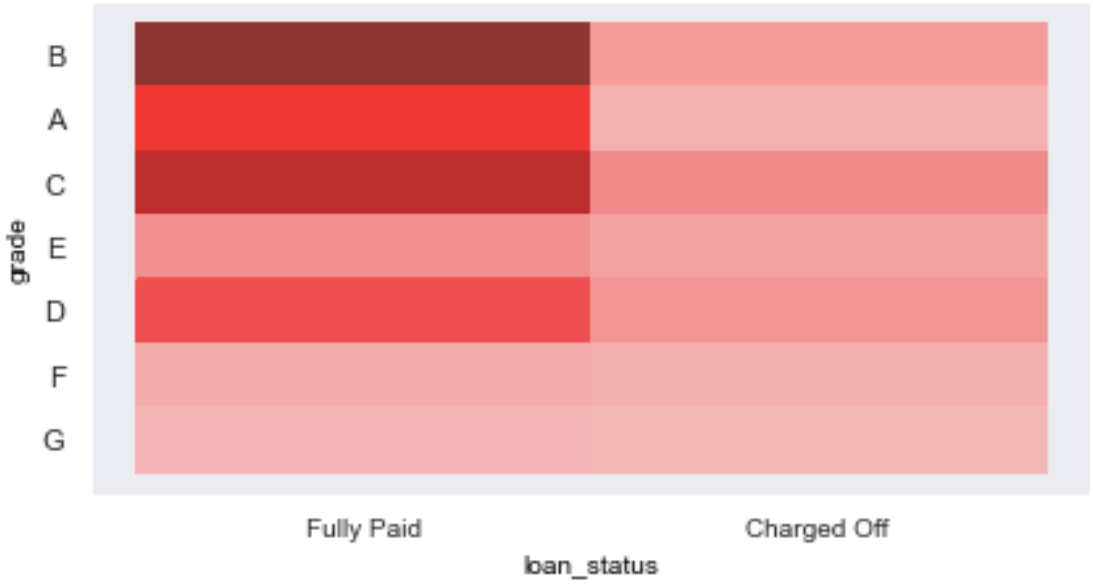
The majority of people have home ownership as :

Majority of the customers who applied for loan had Mortgaged their house.
'home_ownership' is high for 'MORTGAGE'



People with grades ‘A’ are more likely to fully pay their loan.

From the heatmap and the crosstab below it is evident that customers with grade 'A' are high likely to pay the full loan amount.
Customers with Grade A are more likely to pay full loam amount than other grades.
So, it is True that grade A are more likely to pay their loan.

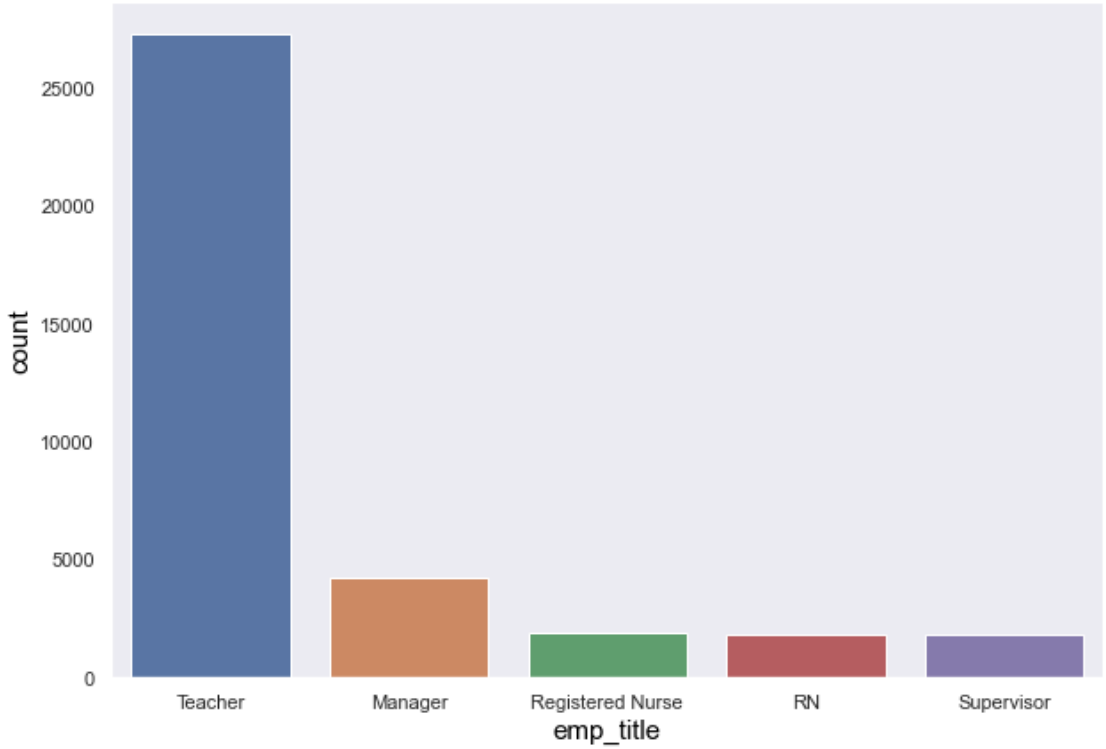


:

grade	A	B	C	D	E	F	G	Fraction
loan_status								
Charged Off	0.06	0.12	0.21	0.28	0.37	0.43	0.48	0.19
Fully Paid	0.94	0.88	0.79	0.72	0.63	0.57	0.52	0.81

Name the top 2 afforded job titles.

Teacher and Manager are the top 2 employee titles of the customers.



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

F1 score is considered as the best metric in this case as the bank should not loose genuine customers or should not go into financial crisis. Here F1 score gives the harmonic mean of precision and recall.

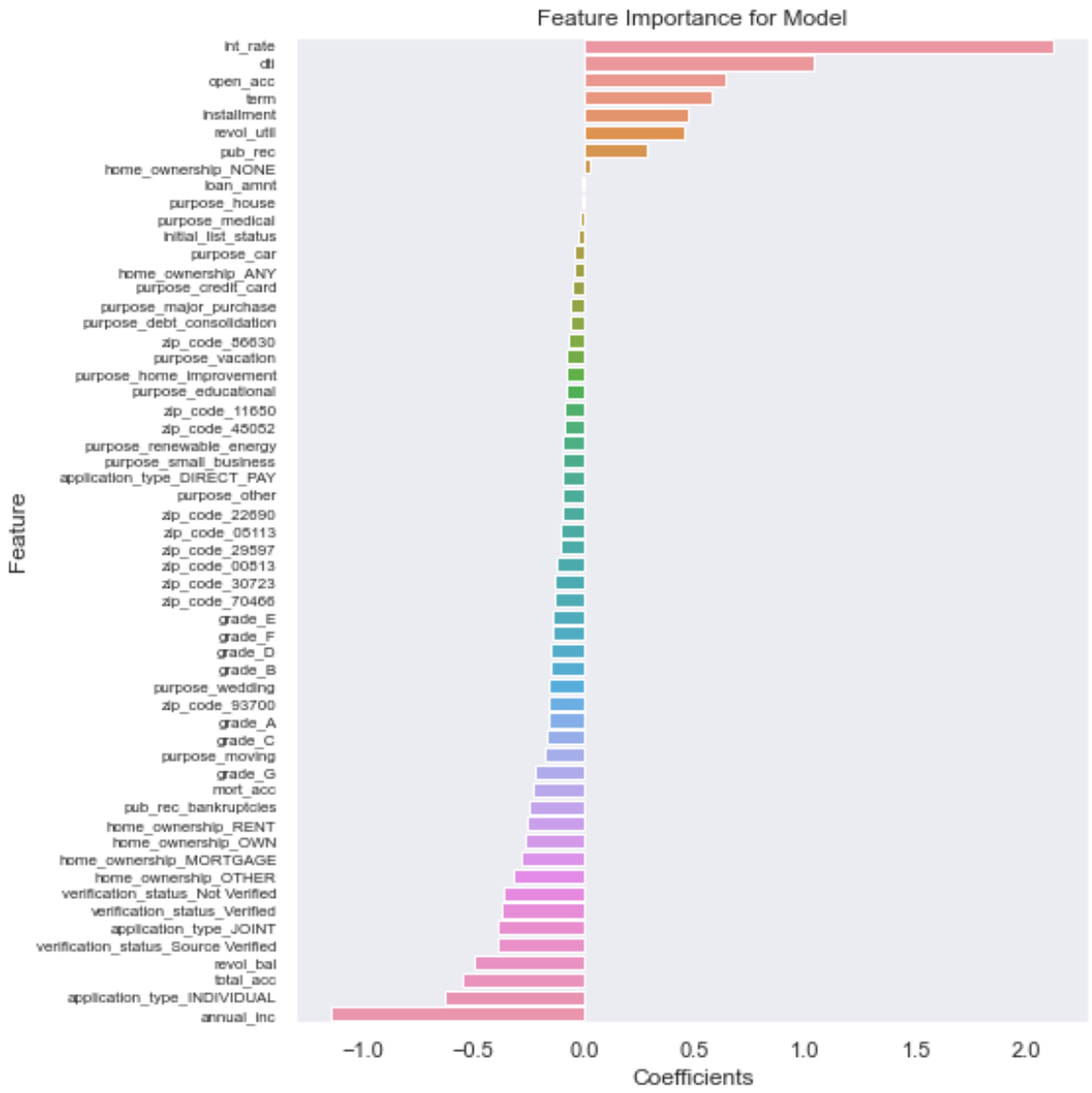
How does the gap in precision and recall affect the bank?

From the below image it is observed as precision value is 0.82 and recall value is 0.98.
If Recall value is low (i.e. FN are high), means customers who actually should be charged off are considered as fully paid customers.
where bank is loosing the huge amount of money(quantity).
If Precision value is low (i.e. FP are high), means customers who paid the loan fully are considered as charged off which indicates bank loosing genuine customers.It decreases the number of quality customers of the bank.

	precision	recall	f1-score	support
0.0	0.82	0.98	0.89	48388
1.0	0.51	0.07	0.12	11399
accuracy			0.81	59787
macro avg	0.66	0.53	0.51	59787
weighted avg	0.76	0.81	0.75	59787

Which were the features that heavily affected the outcome?

'int_rate', 'dti' ,'open_acc','term','annual_income' are the features that heavily affects the outcome.



Will the results be affected by geographical location?
Yes, zip_code derived from address has significant impact on the outcome.

RECOMENDATIONS:

FOR BETTER MODEL BUILDING:

Class Imbalance:

- **Cost-sensitive learning:** Assigning higher costs to misclassifying loan defaults (class 1) during model training. This encourages the model to prioritize correctly identifying defaults even if it means sacrificing some accuracy on the majority class (fully paid loans).
- **Oversampling or undersampling:** Oversampling the minority class (charged-off loans) or undersampling the majority class (fully paid loans) to

create a more balanced dataset. This can help the model learn more effectively from the underrepresented class.

Improve Model Generalizability:

- **Hyperparameter tuning:** Experimenting with different hyperparameter values for chosen machine learning algorithm to potentially improve its performance. Exploring regularization techniques to prevent overfitting and improve generalizability to unseen data.

Feature Engineering:

- **Feature selection:** Analyzing the correlations between features and identify redundant or irrelevant ones for removal. This can improve model performance and reduce training time.
- **Feature creation:** Creating new features based on existing ones. For example, a debt-to-income ratio could be derived from income and total debt.

FOR LOANTAP :

- Ensuring every customer is verified with clear internal company policy or any trusted third-party to avoid false customers,

By implementing these recommendations and continuously iterating the model development process, we can build a more robust, fair, and generalizable model that effectively assists LoanTap in their loan underwriting process.