Bank Loan Pricing

Analyzing the Optimal Interest Rate for Personal Loans and Customer Segmentation.

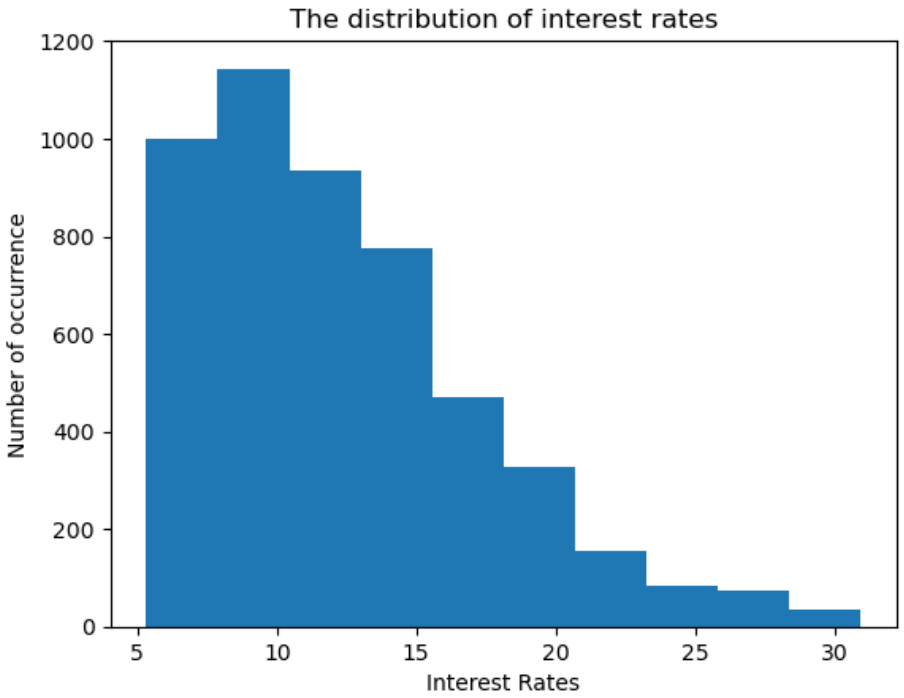
This Photo b Unknown Author is

SREOSHI CHOWDHURI

Project Objective

- Determining the optimal interest rate for the personal loans given by the bank on the basis of the different features of the loan dataset using supervised learning methods.
- It is essential for banks to maintain competitiveness, manage risk, ensure profitability, comply with regulations and meet customer expectations.
- Secondly, segregating the customer base into different clusters and giving marketing and credit recommendations based on the customer's entire banking relationship using unsupervised learning methods.
- This in turn would help to raise sales and boost profitability.

Target Variable : Interest Rate

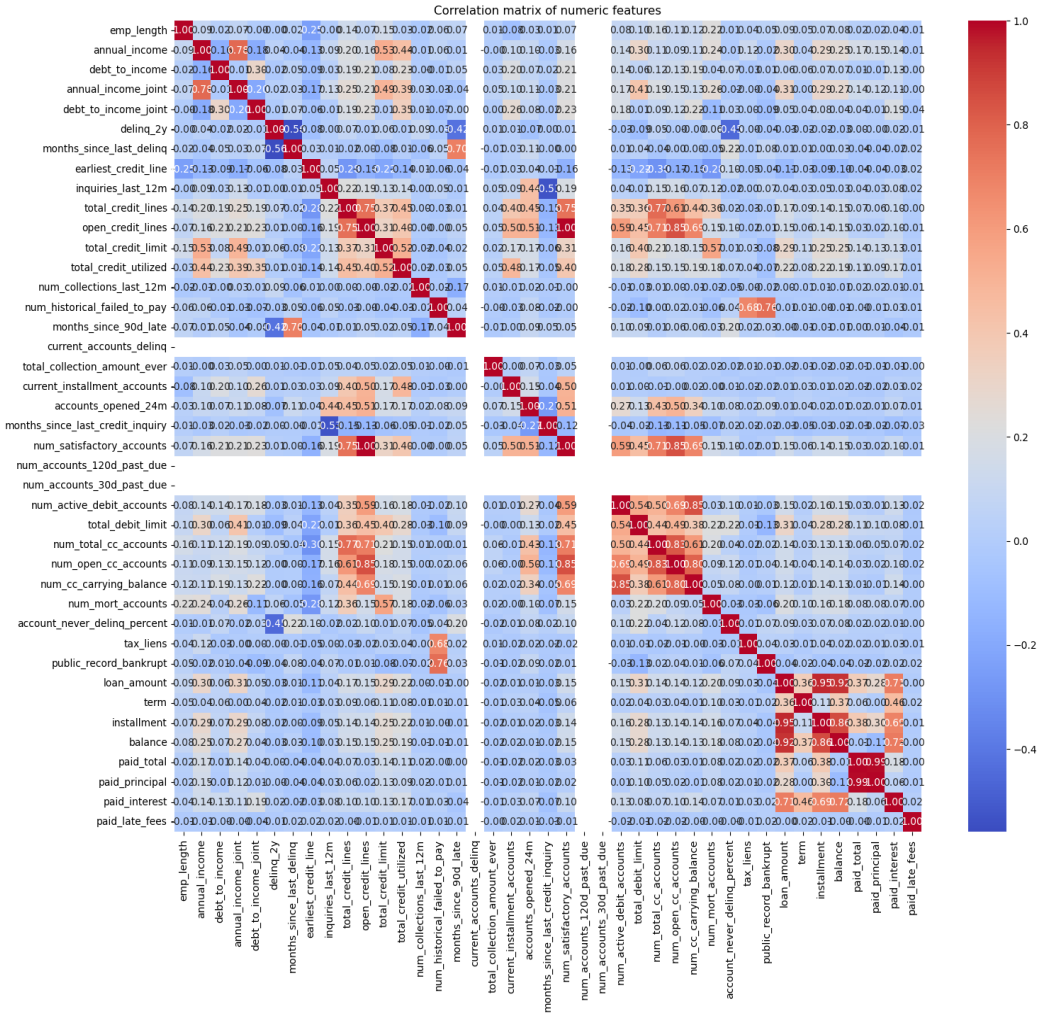


We see that the interest rate distribution is positively skewed which suggests that some interest rates in the dataset are significantly higher than the average or median rate.
From a Risk perspective, a non-normal distribution might suggest that some borrowers are subject to higher interest rates due to perceived credit risk and market conditions.

Correlation of the top 6 numeric features with interest rate.

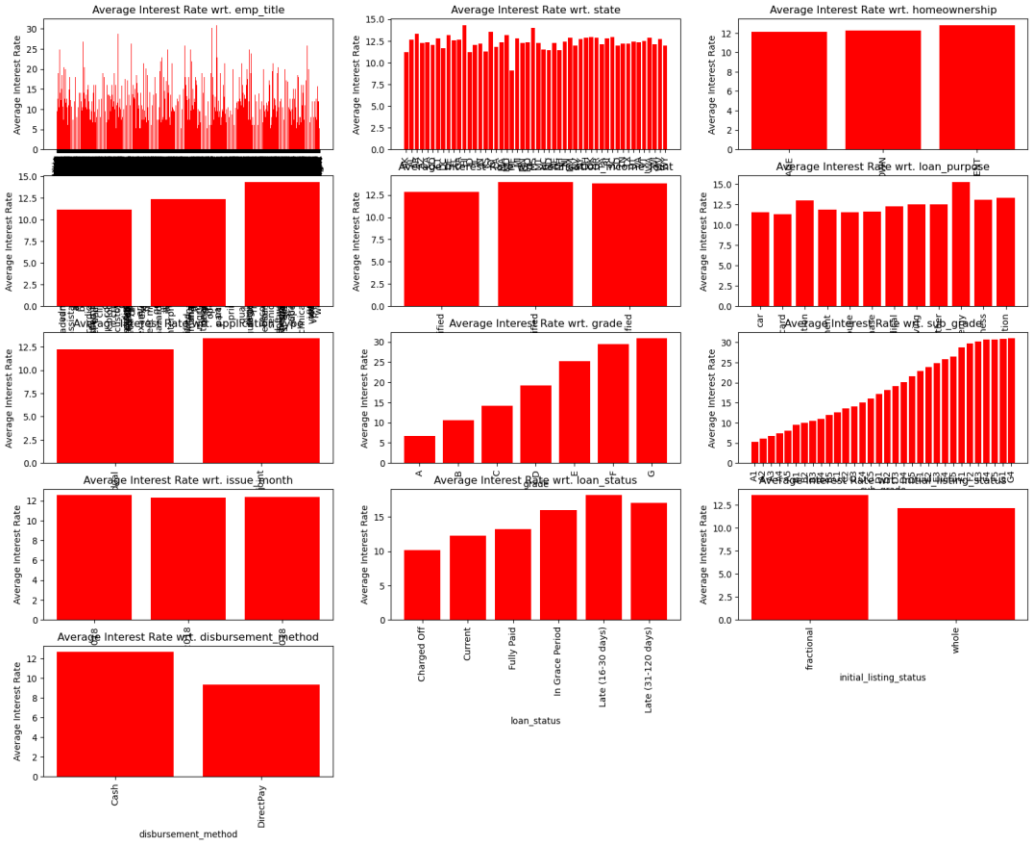
Feature	Correlation
Paid_interest	0.517961
Term	0.362682
Debt_to_income_joint	0.283040
Total_debit_limit	0.250370
Annual_income_joint	0.173321
Debt_to_income	0.136985

Correlation Matrix



Categorical Features

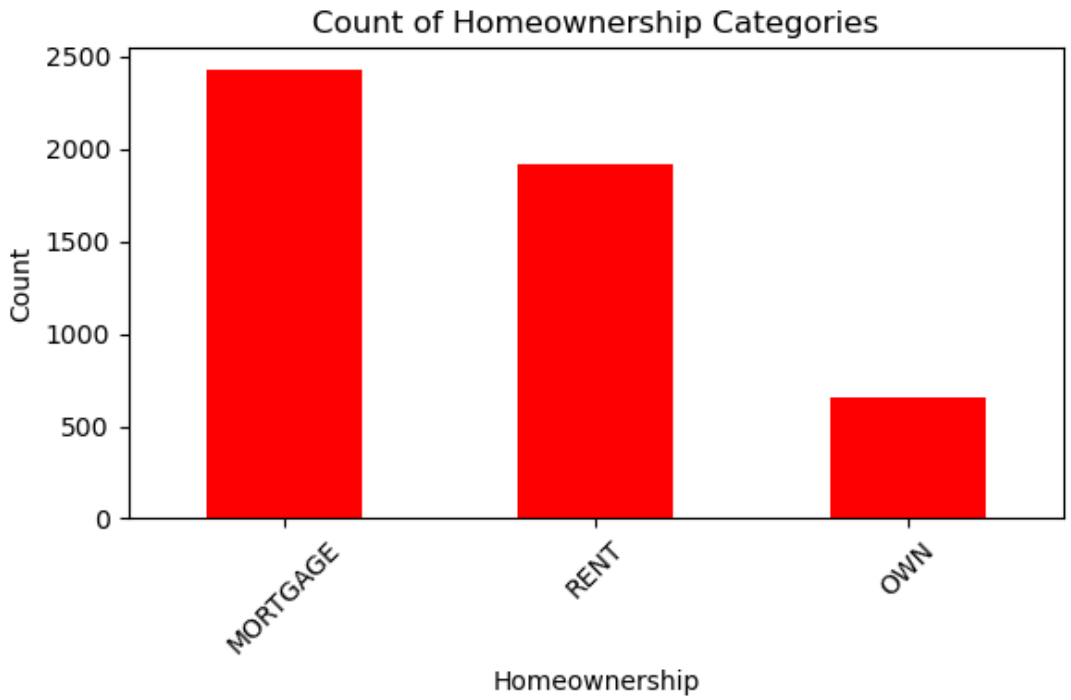
Exploratory data analysis between the target variable and the categorical features:
'Grade' is the only feature which is used in the regression model.



Home Ownership Segmentation

Customers with high annual_income_joint (annual_inc_joint >Mean annual_inc_jnt) and:

a) Rent: Can be pitched special home loan offers.
b) Mortgage: If their mortgage is with a competing lender, our bank can provide lucrative refinance rates to capture this customer segment.
c) Own: Can be approached for a mortgage loan for a second house as investment property.



Regression Models

Linear Regression

It is a predictive analysis approach which models the relationship between a dependent variable(target) and one or more independent variables using a linear equation. It is simple to implement, interpret and efficient to train.

Lasso Regression

Lasso (Least Absolute Shrinkage and Selection Operator) is a regression analysis method that performs both variable selection and regularization to improve the prediction accuracy and interpretability of the statistical model. It is a regularization method that adds a penalty term(i.e. absolute sum of the coefficients) to the OLS objective function. Helps in feature selection by shrinking the coefficients of less important features to exactly zero.

Ridge Regression

It is a regularization technique where a penalty term is added to the OLS objective function. This penalty term is proportional to the sum of the squares of the coefficients (L2 regularization). The objective of Ridge regression is to shrink the coefficients towards zero. This helps in reducing the model complexity and mitigating the problem of multicollinearity, where predictor variables are highly correlated.

Regression Models

Elastic Net Regression

Elastic Net regression is a regularization technique that combines both L1 (Lasso) and L2 (Ridge) penalties in its objective function. Elastic Net strikes a balance between Ridge and Lasso by including both penalties, which can be advantageous when both types of regularization are needed. Like Lasso, Elastic Net can perform feature selection by shrinking some coefficients to zero, effectively ignoring less important predictors. Like Ridge, Elastic Net can handle multicollinearity well because of the L2 penalty.

Random Forest Regression

Random Forest is an ensemble of decision trees, where each tree is trained on a random subset of the training data and a random subset of the features. In regression tasks, the output of the Random Forest is the average (or sometimes the median) prediction of the individual trees. Random Forests are less prone to overfitting compared to individual decision trees, especially when the number of trees (ensemble size) is large. They can capture non-linear relationships between features and the target variable. Random Forests can provide insights into which features are most important for prediction.

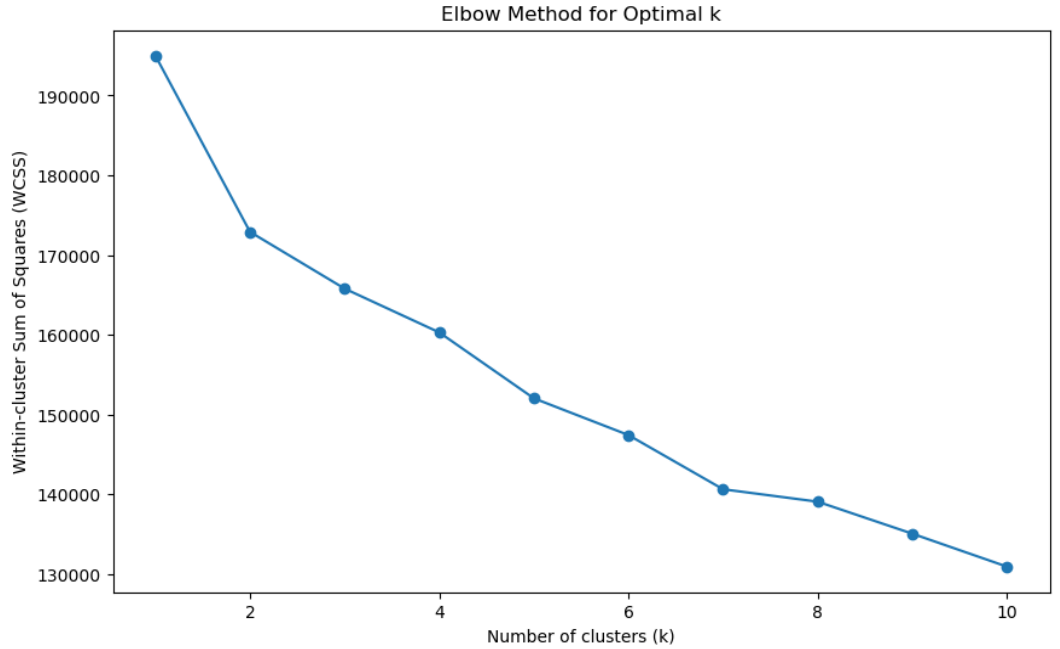
Regression Model Comparison Table

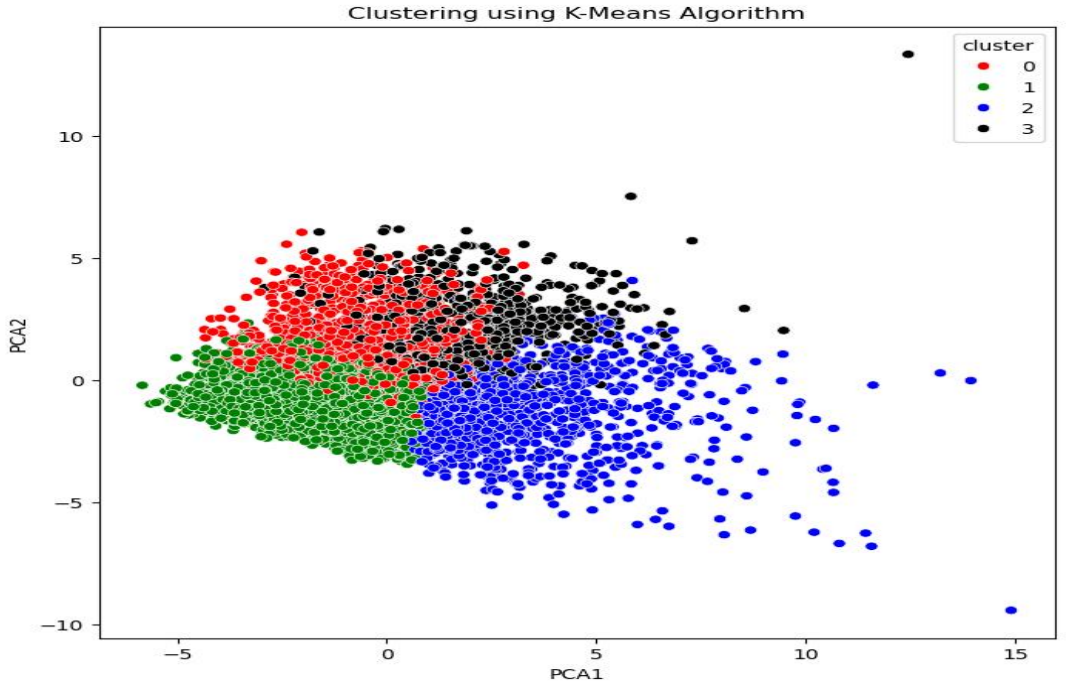
Model	R-Sq. Training	R-Sq. Test	MAE	MSE	RMSE	MAPE
Linear Regression	0.9440935	0.938025	0.07097	0.008393	0.09161	2.98065
Lasso Regression	0.94409204	0.938048	0.00839	0.00839	0.091601	2.9796
Ridge Regression	0.9440935267	0.9380257	0.070972	0.008393	0.091618	2.980657
Elastic Net Regression	0.9440518	0.938074	0.0709022	0.008387	0.09158	2.97761
Random Forest Regression	0.9912213	0.9338006	0.07	0.01	0.09	3.09

Best Model for predicting the interest rate based on the different features:
It was a tough choice selecting the best model, especially between Lasso and Elastic Net Regression but overall, I feel Elastic Net Regression is the best model we can select for predicting the loan interest rate based on the different features since R-squared Test for Elastic Net Regression is slightly higher than Lasso and the errors are more or less the same (MAE for Lasso is although slightly lower than Elastic Regression).

K-Means Clustering:

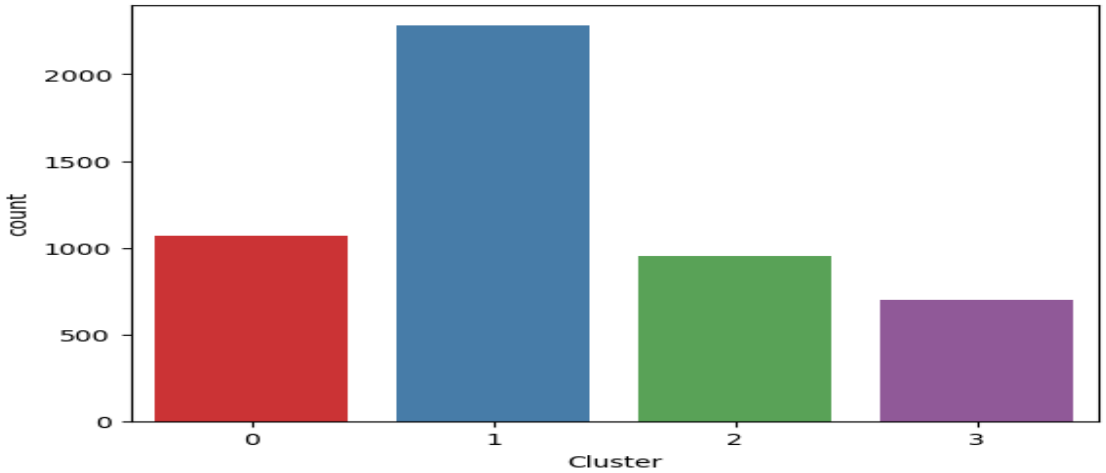
This is a powerful technique which is used for customer segmentation, enabling banks to uncover actionable insights from customer data & tailor strategies to specific customer segments effectively.
The Elbow Method is used to determine the optimal number of clusters.
From the diagram it's visible that the customers have been segregated into 4 clusters.

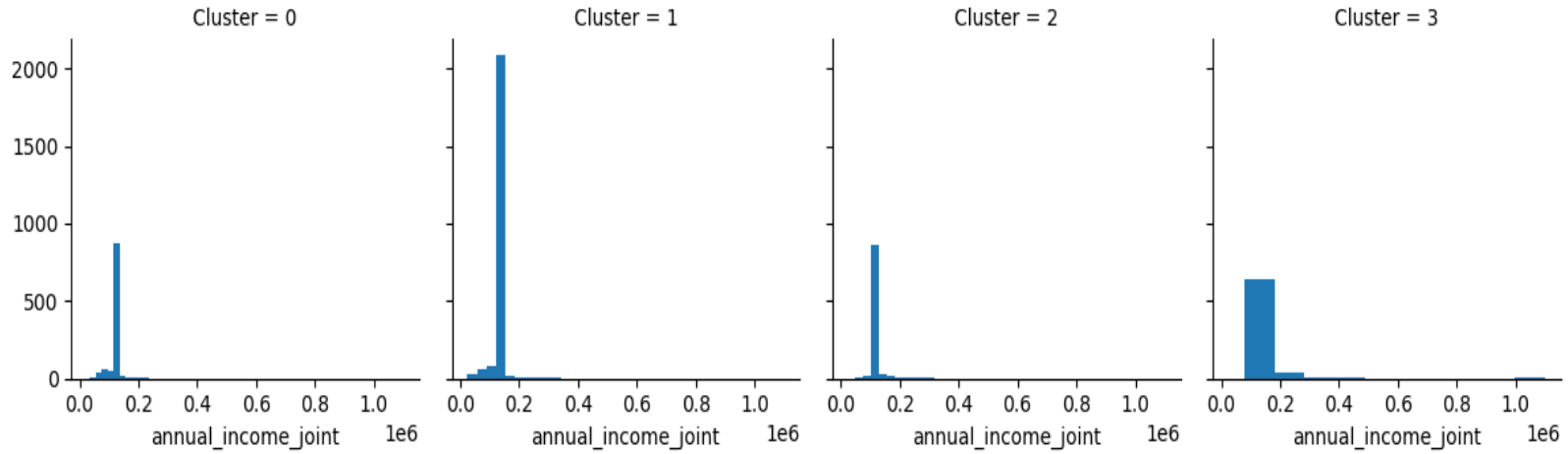




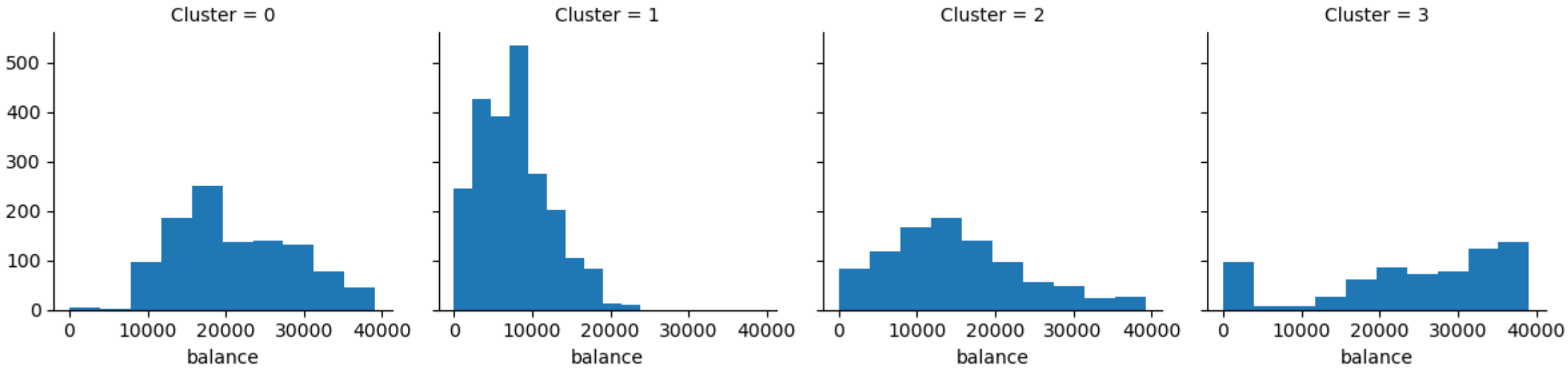
The Clusters

Customers in Cluster 3 seem to have higher annual income joint hence we can discuss savings and wealth management options with them.





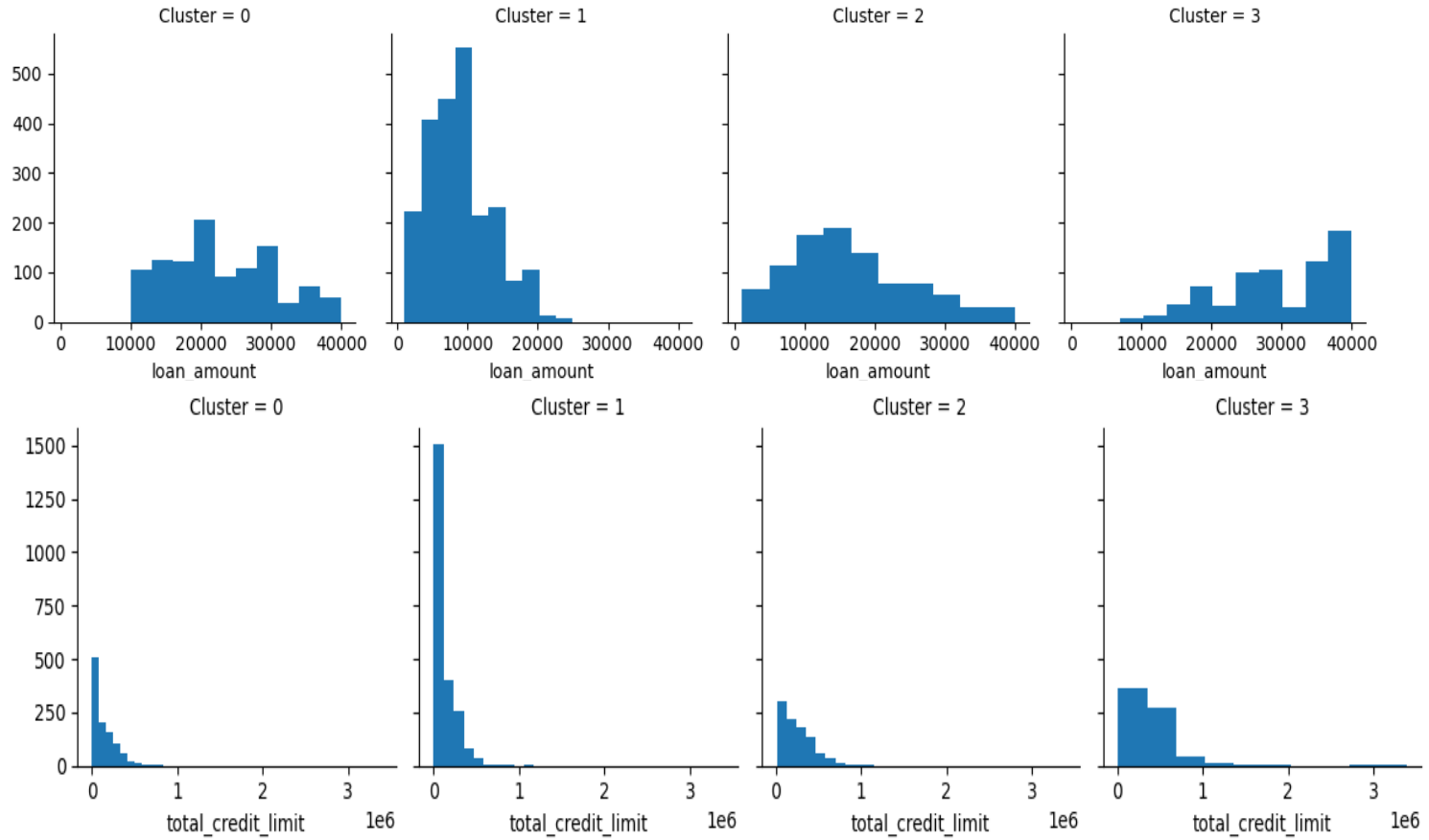
The Clusters



The Balance kept by the customers in cluster 3 is also higher compared to the other groups & hence we can recommend them to upgrade to the bank's priority service i.e., Private Client Account which requires a higher balance relationship with the bank.

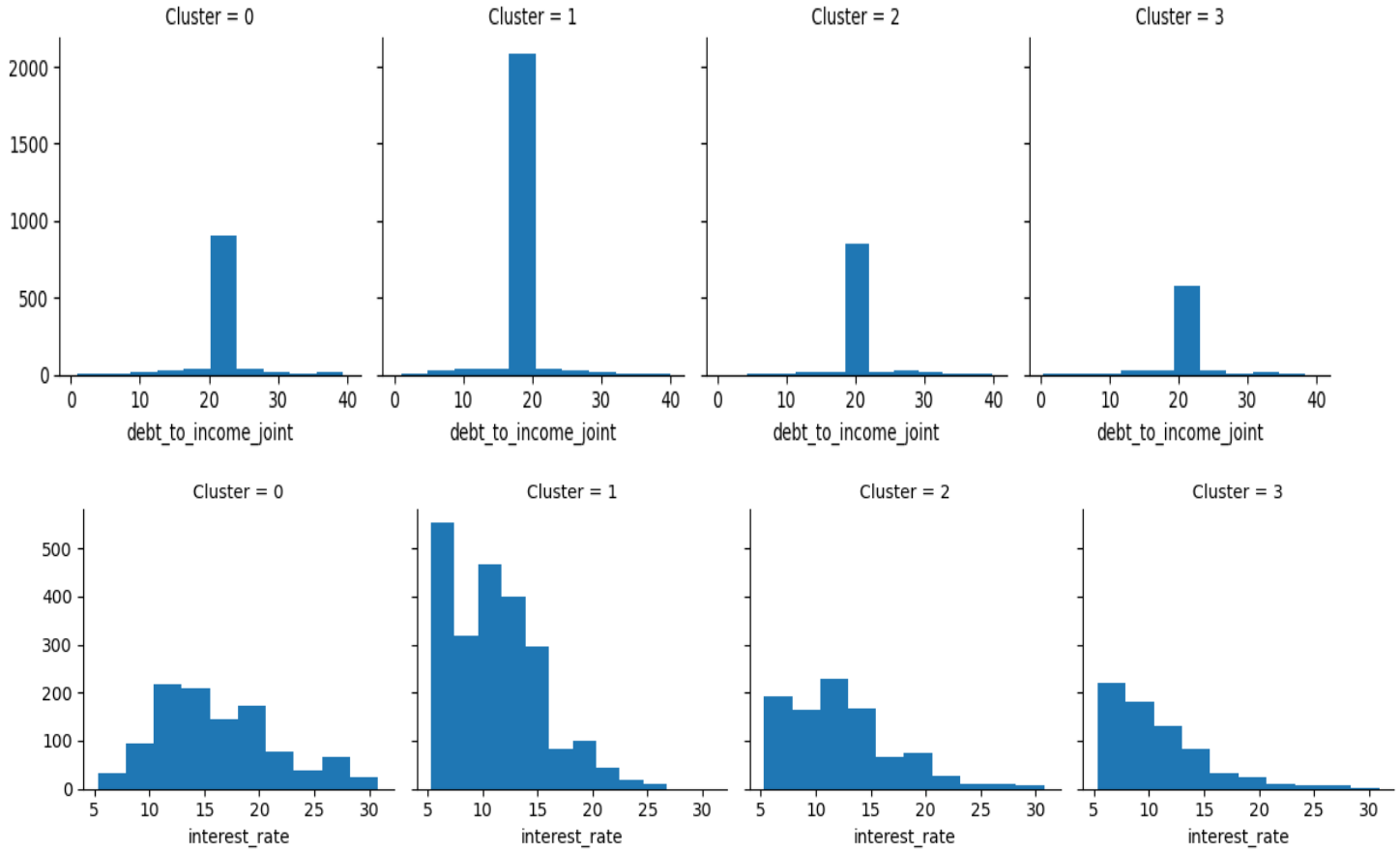
The Clusters

We see that the total_credit_limit is the highest for cluster 3 since they have high income and maintain high balance. Loan Amount given is also higher for Cluster 3 (since as security the bank has high balance for them & high income).

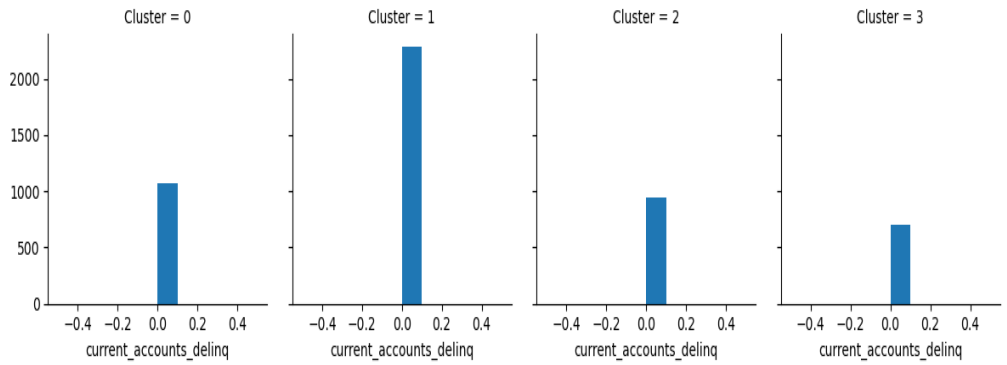


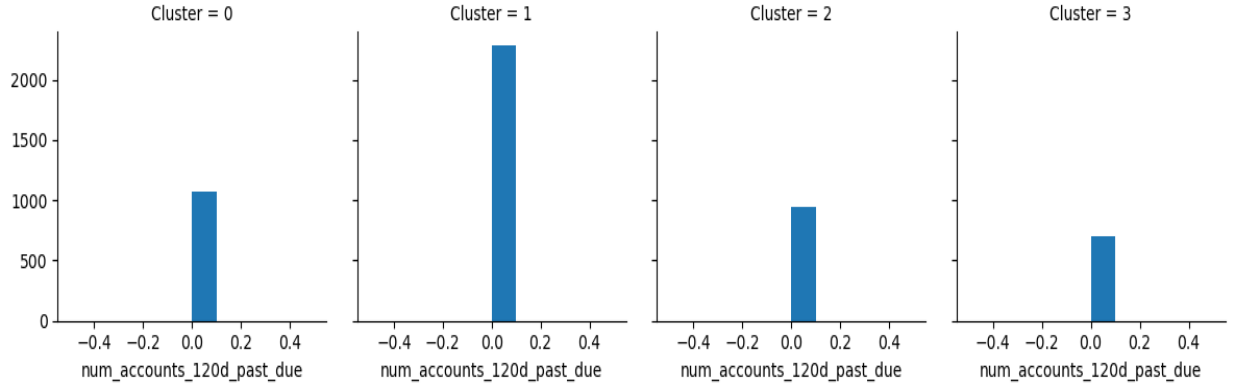
The Clusters

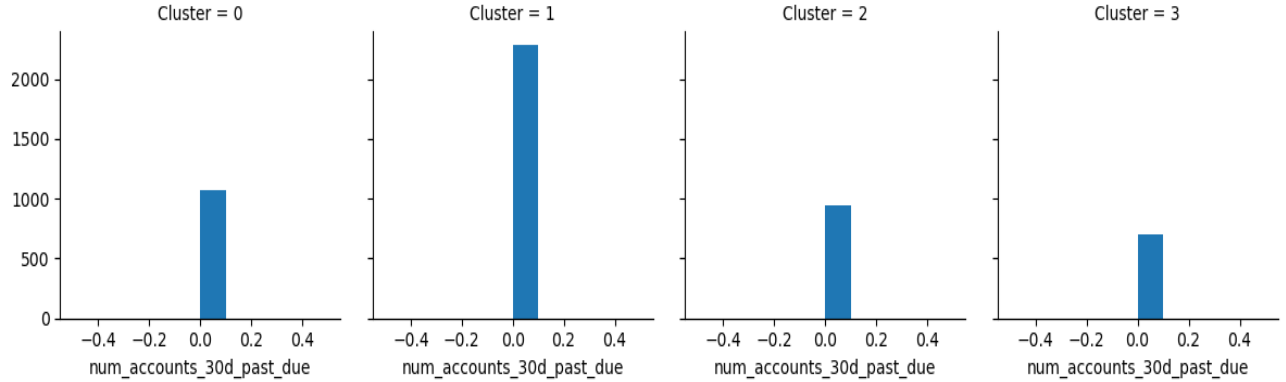
Cluster 0 has higher interest rate as the debt_to_income joint ratio is higher for cluster 0 (Higher the debt_to_income_ratio, riskier is the loan & hence it is priced higher).



The Clusters

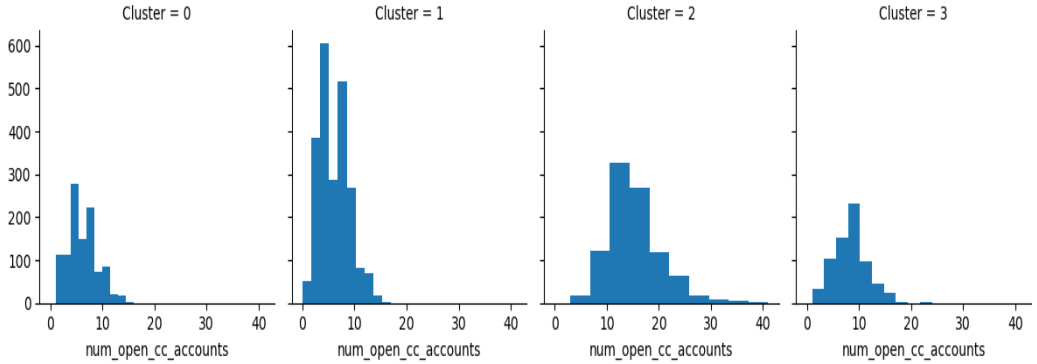


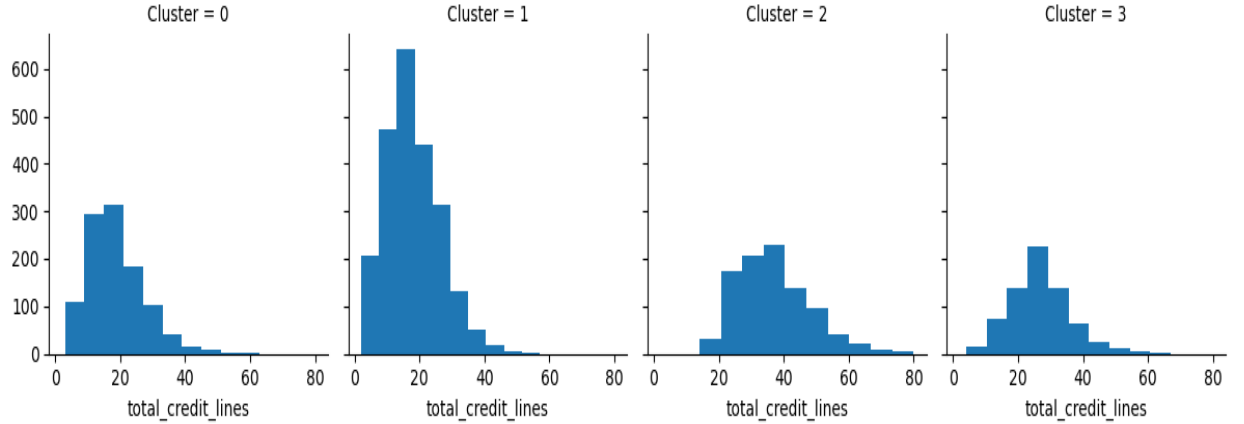


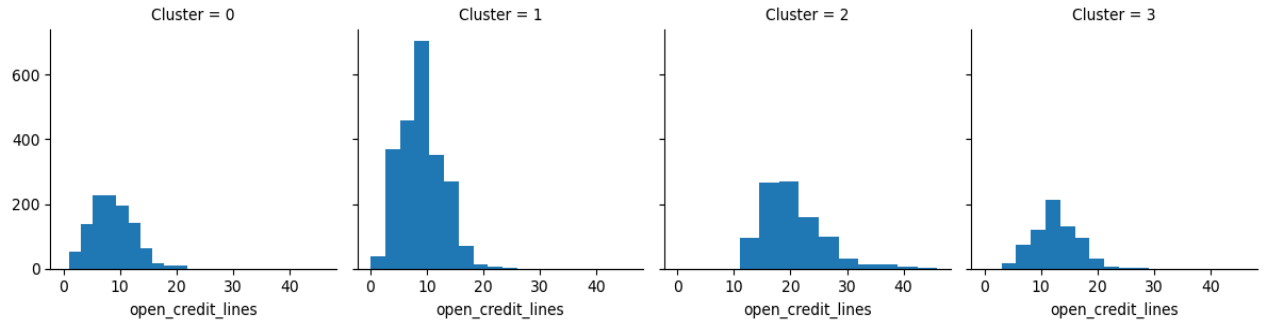


We should be very careful before giving any more loans to cluster 1 and should judge their eligibility very carefully as num_accounts_120day_past_due, num_accounts_30d_past_due,current_accounts_delinq are highest for cluster 1.

The Clusters







Customers with highest total credit line, open credit line, num_open_creditcard_accounts fall in cluster 2 hence we should judge very carefully their eligibility before providing any more loans.

Conclusion

- Elastic Net Regression is the best prediction model in this case for determining the optimal loan interest rate based on the important features.
- Determining the optimal interest rate ensures that the bank earns a competitive return on it's lending activities & is also able to manage credit risk.
- The customer segmentation into different clusters help in providing different promotional offers and credit decisioning which in turn help to boost sales and raise the bank's profit margin.