Report Project 4 – AI for Finance

Zanolo, Luca

238838

Contents

Evaluation of Missing Data in the Dataset

This section presents the results of the evaluation conducted on the given dataset to assess the presence of missing values and their distribution across different columns.

Upon analyzing the dataset, was identified that there are a total of 3,943 rows containing missing values and the following table displays the count of missing values for each column in the dataset.

Column	Missing Values Count
person_age	0
person_income	0
person_home_ownership	0
person_emp_length	895
loan_intent	0
loan_grade	0
loan_amnt	0
loan_int_rate	3,116
loan_status	0
loan_percent_income	0
cb_person_default_on_file	0
cb_person_cred_hist_length	0

As observed, most columns in the dataset do not contain any missing values. However, two columns, namely person_emp_length and loan_int_rate, exhibit missing values. The column person_emp_length has 895 missing values, while the column loan_int_rate has a significantly higher count of 3,116 missing values.

The missing values in the person_emp_length column represent the employment length of the individuals associated with the loan applications. The absence of this information may hinder the model's ability to capture the relationship between employment length and loan default, which could potentially impact the model's predictive performance.

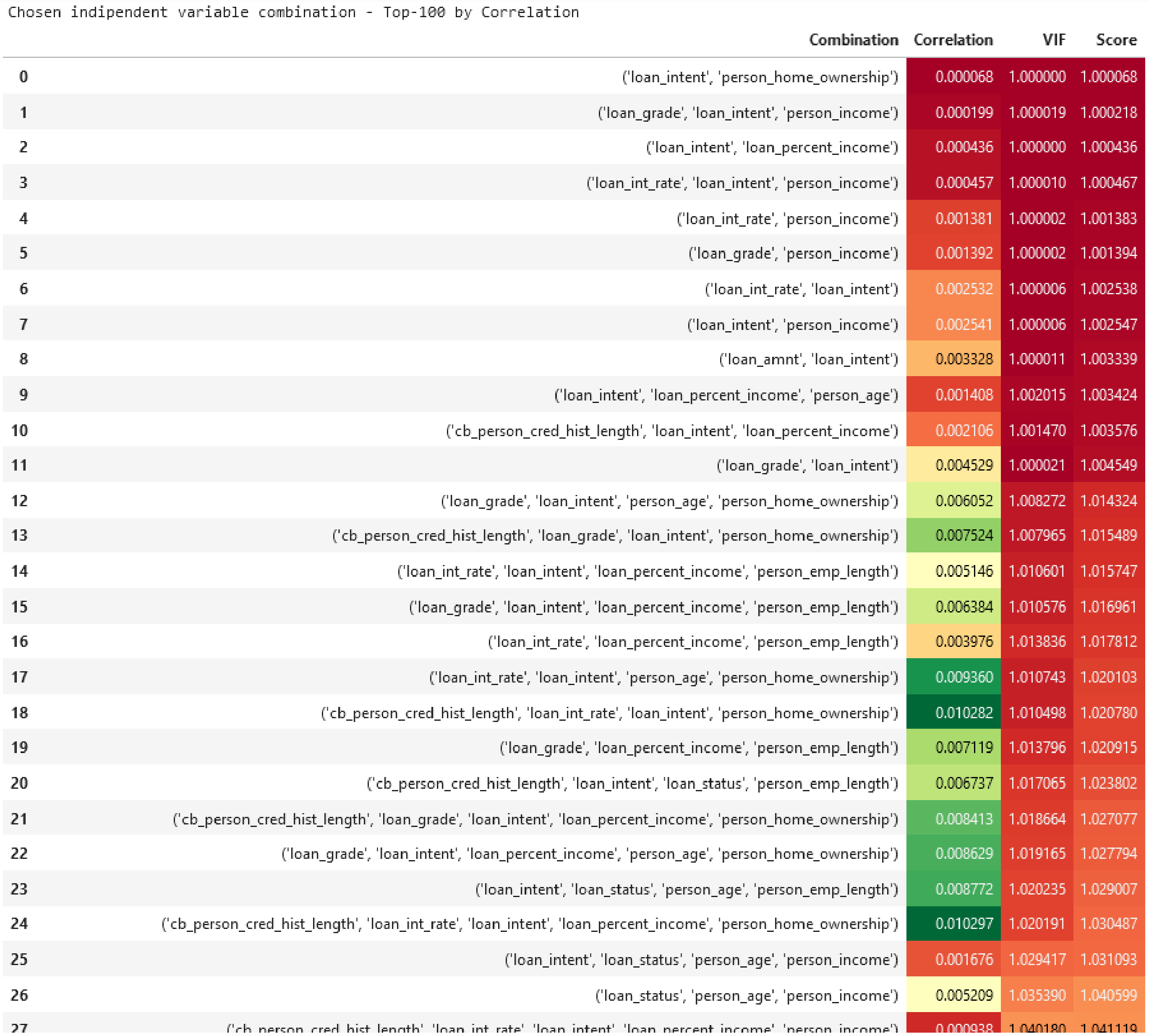
In the subsequent sections of this report, the analysis and modeling will be performed using the modified dataset without the rows that have missing values.

Multicollinearity Analysis

This section presents the results of the multicollinearity analysis conducted on the dataset to identify potential correlations and assess the variance inflation factor (VIF) for different combinations of independent variables.

The analysis considered combinations of independent variables, excluding single-element combinations. The correlation for combinations involving more than two columns was computed as the average of the correlation matrix for the columns involved. Additionally, the VIF was calculated to measure the extent of multicollinearity for each combination.

Based on the analysis, the top 100 combinations, ordered by the "Score," are as follows (only 26 reported):



```
Chosen indipendent variable combination - Top-100 by Correlation
```

	Combination	Correlation	VIF	Score
0	('loan_intent', 'person_home_ownership')	0.000068	1.000000	1.000068
1	('loan_grade', 'loan_intent', 'person_income')	0.000199	1.000019	1.000218
2	('loan_intent', 'loan_percent_income')	0.000436	1.000000	1.000436
3	('loan_int_rate', 'loan_intent', 'person_income')	0.000457	1.000010	1.000467
4	('loan_int_rate', 'person_income')	0.001381	1.000002	1.001383
5	('loan_grade', 'person_income')	0.001392	1.000002	1.001394
6	('loan_int_rate', 'loan_intent')	0.002532	1.000006	1.002538
7	('loan_intent', 'person_income')	0.002541	1.000006	1.002547
8	('loan_amnt', 'loan_intent')	0.003328	1.000011	1.003339
9	('loan_intent', 'loan_percent_income', 'person_age')	0.001408	1.002015	1.003424
10	('cb_person_cred_hist_length', 'loan_intent', 'loan_percent_income')	0.002106	1.001470	1.003576
11	('loan_grade', 'loan_intent')	0.004529	1.000021	1.004549
12	('loan_grade', 'loan_intent', 'person_age', 'person_home_ownership')	0.006052	1.008272	1.014324
13	('cb_person_cred_hist_length', 'loan_grade', 'loan_intent', 'person_home_ownership')	0.007524	1.007965	1.015489
14	('loan_int_rate', 'loan_intent', 'loan_percent_income', 'person_emp_length')	0.005146	1.010601	1.015747
15	('loan_grade', 'loan_intent', 'loan_percent_income', 'person_emp_length')	0.006384	1.010576	1.016961
16	('loan_int_rate', 'loan_percent_income', 'person_emp_length')	0.003976	1.013836	1.017812
17	('loan_int_rate', 'loan_intent', 'person_age', 'person_home_ownership')	0.009360	1.010743	1.020103
18	('cb_person_cred_hist_length', 'loan_int_rate', 'loan_intent', 'person_home_ownership')	0.010282	1.010498	1.020780
19	('loan_grade', 'loan_percent_income', 'person_emp_length')	0.007119	1.013796	1.020915
20	('cb_person_cred_hist_length', 'loan_intent', 'loan_status', 'person_emp_length')	0.006737	1.017065	1.023802
21	('cb_person_cred_hist_length', 'loan_grade', 'loan_intent', 'loan_percent_income', 'person_home_ownership')	0.008413	1.018664	1.027077
22	('loan_grade', 'loan_intent', 'loan_percent_income', 'person_age', 'person_home_ownership')	0.008629	1.019165	1.027794
23	('loan_intent', 'loan_status', 'person_age', 'person_emp_length')	0.008772	1.020235	1.029007
24	('cb_person_cred_hist_length', 'loan_int_rate', 'loan_intent', 'loan_percent_income', 'person_home_ownership')	0.010297	1.020191	1.030487
25	('loan_intent', 'loan_status', 'person_age', 'person_income')	0.001676	1.029417	1.031093
26	('loan_status', 'person_age', 'person_income')	0.005209	1.035390	1.040599
27	('cb_person_cred_hist_length', 'loan_int_rate', 'loan_intent', 'loan_percent_income', 'person_income')	0.000938	1.040180	1.041119

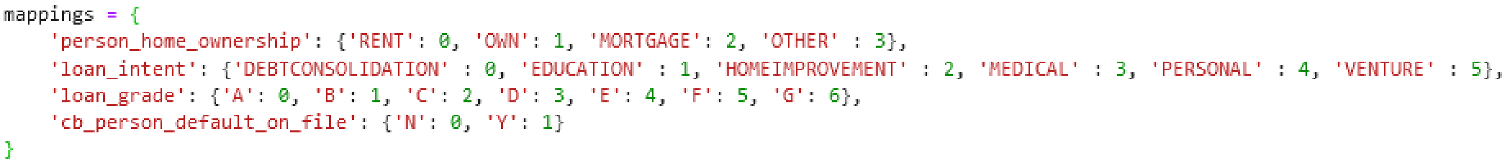
The results reveal that many combinations exhibit low correlation and VIF values, indicating a lack of substantial multicollinearity. However, some combinations show higher correlation and VIF values, suggesting potential multicollinearity concerns.

In the subsequent sections of this report, all the predefined models will be tested on each combination within the top N of the ranking. The objective is to evaluate the performance of these models and assess their ability to accurately predict loan default. By testing the models on various combinations, the goal is to identify the most effective set of independent variables that can provide reliable predictions while mitigating the impact of multicollinearity.

Data Preparation

In the data preparation phase, the following operations were applied to the dataset:

- **Applying Mapping**: To handle nominal categorical columns, a mapping dictionary was defined. This dictionary assigns numerical values to the categorical variables for further analysis. The apply_mappings function was used to replace the categorical values with their corresponding numerical mappings. This ensures consistency and enables mathematical operations on the data.



```
mappings = {
    'person_home_ownership': {'RENT': 0, 'OWN': 1, 'MORTGAGE': 2, 'OTHER' : 3},
    'loan_intent': {'DEBTCONSOLIDATION' : 0, 'EDUCATION' : 1, 'HOMEIMPROVEMENT' : 2, 'MEDICAL' : 3, 'PERSONAL' : 4, 'VENTURE' : 5},
    'loan_grade': {'A': 0, 'B': 1, 'C': 2, 'D': 3, 'E': 4, 'F': 5, 'G': 6},
    'cb_person_default_on_file': {'N': 0, 'Y': 1}
}
```

- **Standardizing Data**: The StandardScaler class from the scikit-learn library was used to standardize columns in the dataset.

The overall data preparation process ensures that missing values are handled, categorical variables are appropriately encoded, and numerical features are standardized.

Analysis

This chapter focuses on the application of different models for predicting loan defaults. The following models were utilized in this analysis:

- *Logistic Regression (Logit)*
- *K-Nearest Neighbors (KNN)*
- *Gradient Boosting*
- *Random Forest*
- *Neural Network*

To evaluate the performance of these models, an out-of-sample analysis was conducted. The dataset was divided into training and testing sets, with 33% of the data reserved for testing and the remaining portion used for training.

Independent Variable Combinations

In the analysis, various independent variable combinations were evaluated to determine their impact on the performance of the models. The following combinations stood out as having the best average performance across the models.

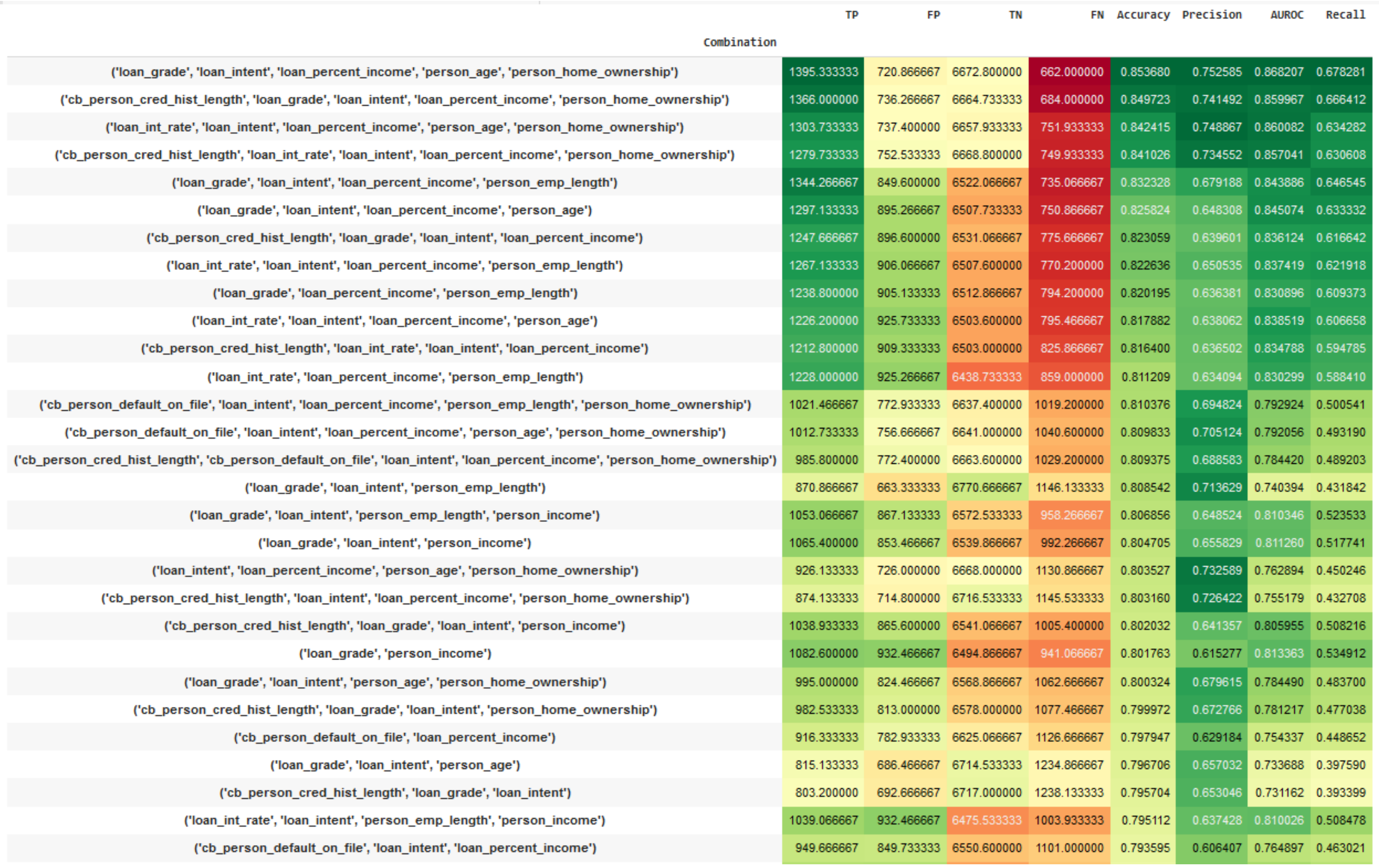
The table presents a scoreboard of the models' performances with their corresponding independent variable combinations. While we won't delve into the details of each combination, let's briefly discuss some of the best-performing combinations.

One notable combination is (loan_grade, loan_intent, loan_percent_income, person_age, person_home_ownership). This combination achieves an accuracy of 0.8537, indicating that approximately 85.37% of predictions are correct. The precision value of 0.7526 suggests that among all the predicted defaults, around 75.26% are truly defaults. Additionally, the AUROC score of 0.8682 indicates good discriminatory power in distinguishing between default and non-default cases, and the recall value of 0.6783 implies that around 67.83% of all actual defaults are captured.

Another strong combination is ('cb_person_cred_hist_length', 'loan_grade', 'loan_intent', 'loan_percent_income', 'person_home_ownership'). This combination exhibits an accuracy of 0.8497 and a precision value of 0.7415. The AUROC score of 0.8600 indicates effective discrimination between default and non-default instances. The recall value of 0.6664 suggests that approximately 66.64% of all actual defaults are captured by this combination.

In conclusion, the analysis of different independent variable combinations reveals the effectiveness of certain variables in predicting loan defaults. The selected combinations consistently demonstrate high accuracy, precision, AUROC scores, and recall values, indicating their strong predictive power.

Variables such as loan grade, loan intent, loan percent income, person age, and person home ownership consistently emerge as key contributors to the prediction of loan defaults. These variables capture important aspects related to the borrower's financial situation, credit history, and loan characteristics.



	TP	FP	TN	FN	Accuracy	Precision	AUROC	Recall
Combination								
('loan_grade', 'loan_intent', 'loan_percent_income', 'person_age', 'person_home_ownership')	1395.333333	720.866667	6672.800000	662.000000	0.853680	0.752585	0.868207	0.678281
('cb_person_cred_hist_length', 'loan_grade', 'loan_intent', 'loan_percent_income', 'person_home_ownership')	1366.000000	736.266667	6664.733333	684.000000	0.849723	0.741492	0.859967	0.666412
('loan_int_rate', 'loan_intent', 'loan_percent_income', 'person_age', 'person_home_ownership')	1303.733333	737.400000	6657.933333	751.933333	0.842415	0.748867	0.860082	0.634282
('cb_person_cred_hist_length', 'loan_int_rate', 'loan_intent', 'loan_percent_income', 'person_home_ownership')	1279.733333	752.533333	6668.800000	749.933333	0.841026	0.734552	0.857041	0.630608
('loan_grade', 'loan_intent', 'loan_percent_income', 'person_emp_length')	1344.266667	849.600000	6522.066667	735.066667	0.832328	0.679188	0.843886	0.646545
('loan_grade', 'loan_intent', 'loan_percent_income', 'person_age')	1297.133333	895.266667	6507.733333	750.866667	0.825824	0.648308	0.845074	0.633332
('cb_person_cred_hist_length', 'loan_grade', 'loan_intent', 'loan_percent_income')	1247.666667	896.600000	6531.066667	775.666667	0.823059	0.639601	0.836124	0.616642
('loan_int_rate', 'loan_intent', 'loan_percent_income', 'person_emp_length')	1267.133333	906.066667	6507.600000	770.200000	0.822636	0.650535	0.837419	0.621918
('loan_grade', 'loan_percent_income', 'person_emp_length')	1238.800000	905.133333	6512.866667	794.200000	0.820195	0.636381	0.830896	0.609373
('loan_int_rate', 'loan_intent', 'loan_percent_income', 'person_age')	1226.200000	925.733333	6503.600000	795.466667	0.817882	0.638062	0.838519	0.606658
('cb_person_cred_hist_length', 'loan_int_rate', 'loan_intent', 'loan_percent_income')	1212.800000	909.333333	6503.000000	825.866667	0.816400	0.636502	0.834788	0.594785
('loan_int_rate', 'loan_percent_income', 'person_emp_length')	1228.000000	925.266667	6438.733333	859.000000	0.811209	0.634094	0.830299	0.588410
('cb_person_default_on_file', 'loan_intent', 'loan_percent_income', 'person_emp_length', 'person_home_ownership')	1021.466667	772.933333	6637.400000	1019.200000	0.810376	0.694824	0.792924	0.500541
('cb_person_default_on_file', 'loan_intent', 'loan_percent_income', 'person_age', 'person_home_ownership')	1012.733333	756.666667	6641.000000	1040.600000	0.809833	0.705124	0.792056	0.493190
('cb_person_cred_hist_length', 'cb_person_default_on_file', 'loan_intent', 'loan_percent_income', 'person_home_ownership')	985.800000	772.400000	6663.600000	1029.200000	0.809375	0.688583	0.784420	0.489203
('loan_grade', 'loan_intent', 'person_emp_length')	870.866667	663.333333	6770.666667	1146.133333	0.808542	0.713629	0.740394	0.431842
('loan_grade', 'loan_intent', 'person_emp_length', 'person_income')	1053.066667	867.133333	6572.533333	958.266667	0.806856	0.648524	0.810346	0.523533
('loan_grade', 'loan_intent', 'person_income')	1065.400000	853.466667	6539.866667	992.266667	0.804705	0.655829	0.811260	0.517741
('loan_intent', 'loan_percent_income', 'person_age', 'person_home_ownership')	926.133333	726.000000	6668.000000	1130.866667	0.803527	0.732589	0.762894	0.450246
('cb_person_cred_hist_length', 'loan_intent', 'loan_percent_income', 'person_home_ownership')	874.133333	714.800000	6716.533333	1145.533333	0.803160	0.726422	0.755179	0.432708
('cb_person_cred_hist_length', 'loan_grade', 'loan_intent', 'person_income')	1038.933333	865.600000	6541.066667	1005.400000	0.802032	0.641357	0.805955	0.508216
('loan_grade', 'person_income')	1082.600000	932.466667	6494.866667	941.066667	0.801763	0.615277	0.813363	0.534912
('loan_grade', 'loan_intent', 'person_age', 'person_home_ownership')	995.000000	824.466667	6568.866667	1062.666667	0.800324	0.679615	0.784490	0.483700
('cb_person_cred_hist_length', 'loan_grade', 'loan_intent', 'person_home_ownership')	982.533333	813.000000	6578.000000	1077.466667	0.799972	0.672766	0.781217	0.477038
('cb_person_default_on_file', 'loan_percent_income')	916.333333	782.933333	6625.066667	1126.666667	0.797947	0.629184	0.754337	0.448652
('loan_grade', 'loan_intent', 'person_age')	815.133333	686.466667	6714.533333	1234.866667	0.796706	0.657032	0.733688	0.397590
('cb_person_cred_hist_length', 'loan_grade', 'loan_intent')	803.200000	692.666667	6717.000000	1238.133333	0.795704	0.653046	0.731162	0.393399
('loan_int_rate', 'loan_intent', 'person_emp_length', 'person_income')	1039.066667	932.466667	6475.533333	1003.933333	0.795112	0.637428	0.810026	0.508478
('cb_person_default_on_file', 'loan_intent', 'loan_percent_income')	949.666667	849.733333	6550.600000	1101.000000	0.793595	0.606407	0.764897	0.463021

Features Importance

This section presents the analysis and results of the loan default prediction models. The models were computed using the following approach: the top 100 combinations based on VIF (Variance Inflation Factor) and correlation were selected. For each selected combination, the five models (Logit, Gradient Boosting, Random Forest, KNN and Neural Network) were trained five times. All statistical measures and feature importance were averaged over these iterations, resulting in a single value for each measure.

Firstly, let's discuss the factors that contribute the most to loan defaults. Feature importance was computed for all models except K-Nearest Neighbors (KNN). The average impact of each variable across all computed measures is as follows:

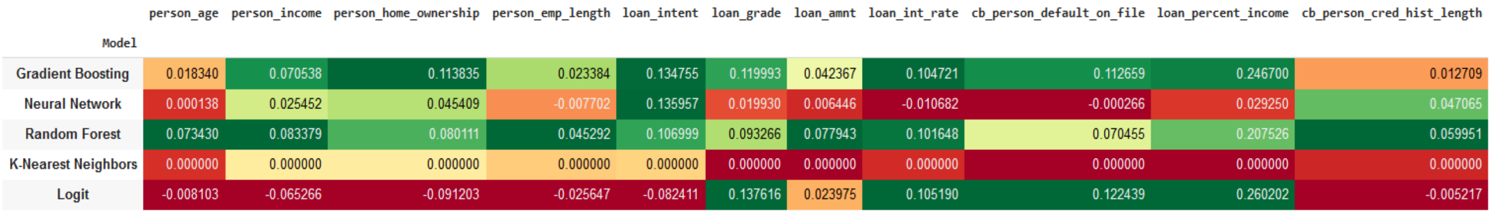
- *person_home_ownership: 0.029630*

- *person_emp_length: 0.007065*
- *cb_person_cred_hist_length: 0.022902*
- *loan_amnt: 0.030146*
- *loan_intent: 0.059060*
- *person_age: 0.016761*
- *loan_percent_income: 0.148736*
- *cb_person_default_on_file: 0.061057*
- *person_income: 0.022820*
- *loan_grade: 0.074161*
- *loan_int_rate: 0.060176*

These values represent the average contribution or impact of each variable in predicting loan defaults across the measures with different models. The loan_percent_income variable stands out as the most influential factor, with an average impact of 0.148736. Other significant contributors include loan_grade (0.074161), cb_person_default_on_file (0.061057), loan_intent (0.059060), and loan_int_rate (0.060176). These variables indicate that a borrower's income in relation to the loan amount, loan grade, credit history, and interest rate significantly impacts the likelihood of loan default. Here are reported more details about these variables:

- **Loan Percent Income:** The ratio of the loan amount to the borrower's income has a high feature importance score. This indicates that borrowers with a higher loan amount relative to their income are more likely to default.
- **Loan Grade:** The loan grade also demonstrates a substantial impact on loan defaults. Loan grade is a measure of the borrower's creditworthiness or risk level assigned by the lender. Lower-grade loans (higher risk) are more prone to default compared to higher-grade loans (lower risk).
- **Credit History:** The length of the borrower's credit history is another important feature. A longer credit history indicates more experience in managing credit, and borrowers with a solid credit history are generally more reliable in meeting their financial obligations.
- **Interest Rate:** The interest rate assigned to the loan also plays a significant role. Higher interest rates can increase the financial load on borrowers, making it more challenging for them to repay the loan on time. Thus, loans with higher interest rates have a higher probability of default.

Now let’s examine the contribution of each variable at a higher granularity by aggregating the data for each model. The table below presents the features contributions for each model in the loan default prediction task.



	person_age	person_income	person_home_ownership	person_emp_length	loan_intent	loan_grade	loan_amnt	loan_int_rate	cb_person_default_on_file	loan_percent_income	cb_person_cred_hist_length
Model											
Gradient Boosting	0.018340	0.070538	0.113835	0.023384	0.134755	0.119993	0.042367	0.104721	0.112659	0.246700	0.012709
Neural Network	0.000138	0.025452	0.045409	-0.007702	0.135957	0.019930	0.006446	-0.010682	-0.000266	0.029250	0.047065
Random Forest	0.073430	0.083379	0.080111	0.045292	0.106999	0.093266	0.077943	0.101648	0.070455	0.207526	0.059951
K-Nearest Neighbors	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
Logit	-0.008103	-0.065266	-0.091203	-0.025647	-0.082411	0.137616	0.023975	0.105190	0.122439	0.260202	-0.005217

Here is the analysis and interpretation of the variable contributions for each model:

- **Gradient Boosting:** Among the variables considered, the most influential factors contributing to loan defaults in the Gradient Boosting model are loan_percent_income (0.2467), loan_intent (0.1348), cb_person_default_on_file (0.1127), and loan_grade (0.12). These variables suggest that the borrower's income in relation to the loan amount, the purpose of the loan, their default history, and loan grade play significant roles in predicting loan defaults.
- **Neural Network:** In the Neural Network model, the variables with the highest impact on loan defaults are loan_intent (0.1359), cb_person_cred_hist_length (0.0471), and person_income (0.0255). This indicates that the

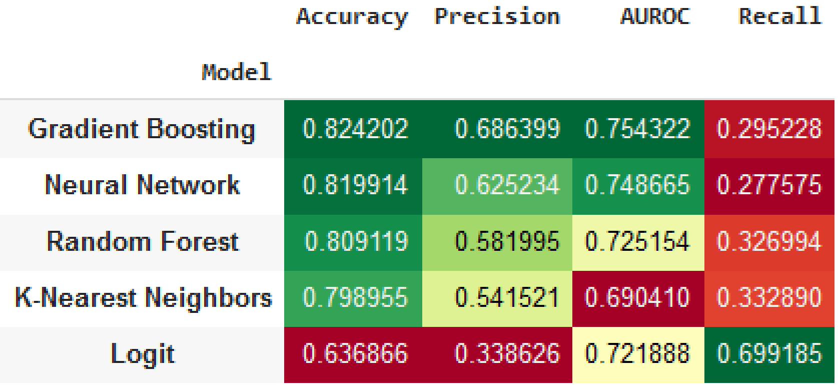
purpose of the loan, the length of the borrower's credit history, and their income level contribute significantly to the prediction of loan defaults in this model.

- **Random Forest:** The Random Forest model identifies loan_percent_income (0.2075), loan_grade (0.0933), and cb_person_default_on_file (0.0705) as the most important variables for predicting loan defaults. These findings align with the previous models, emphasizing the significance of the borrower's income-to-loan ratio, loan grade, and default history.
- **Logit:** In the Logit model, loan_percent_income (0.2602), cb_person_default_on_file (0.1224), and loan_int_rate (0.1052) have the highest impact on loan defaults. These variables highlight the importance of the borrower's income-to-loan ratio, their default history, and the interest rate assigned to the loan.
- K-Nearest Neighbors: The K-Nearest Neighbors model does not provide variable importance measures.

These findings suggest that factors such as the borrower's income relative to the loan amount, loan grade, credit history, default history, and loan purpose play crucial roles in predicting the likelihood of loan defaults.

Models' analysis

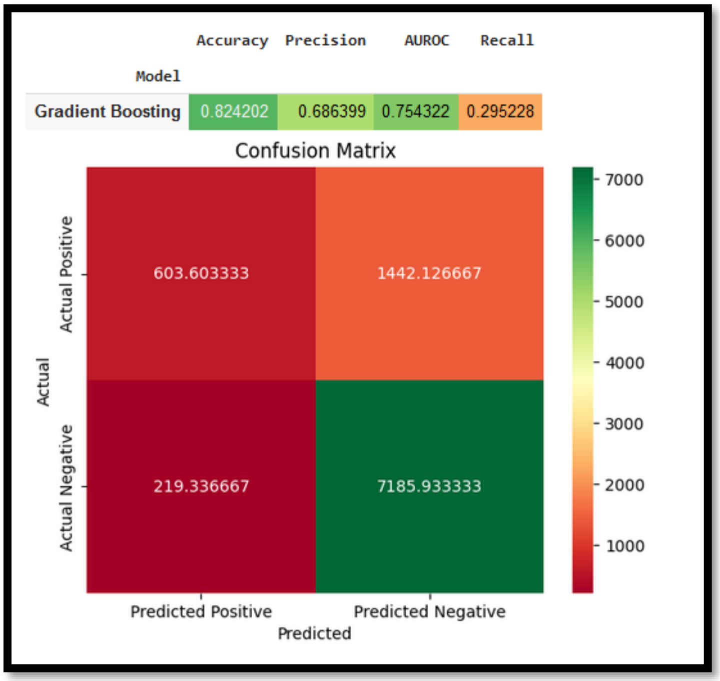
The performance of the loan default prediction models was evaluated using various metrics, including Accuracy, TP (True Positives), FP (False Positives), TN (True Negatives), FN (False Negatives), AUROC, Precision and Recall. The following table provides a summary of the metrics for each of the models use. These values are calculated as an average over all measurements.



	Accuracy	Precision	AUROC	Recall
Model				
Gradient Boosting	0.824202	0.686399	0.754322	0.295228
Neural Network	0.819914	0.625234	0.748665	0.277575
Random Forest	0.809119	0.581995	0.725154	0.326994
K-Nearest Neighbors	0.798955	0.541521	0.690410	0.332890
Logit	0.636866	0.338626	0.721888	0.699185

In comparison, the Gradient Boosting and Neural Network models exhibit relatively higher accuracy, precision, and AUROC scores compared to the other models. These models show better overall performance in predicting loan defaults, while the Random Forest and K-Nearest Neighbors models offer a balance between accuracy and precision. The Logit model performs comparatively poorer, with lower accuracy and precision but a higher recall rate.

Gradient Boosting



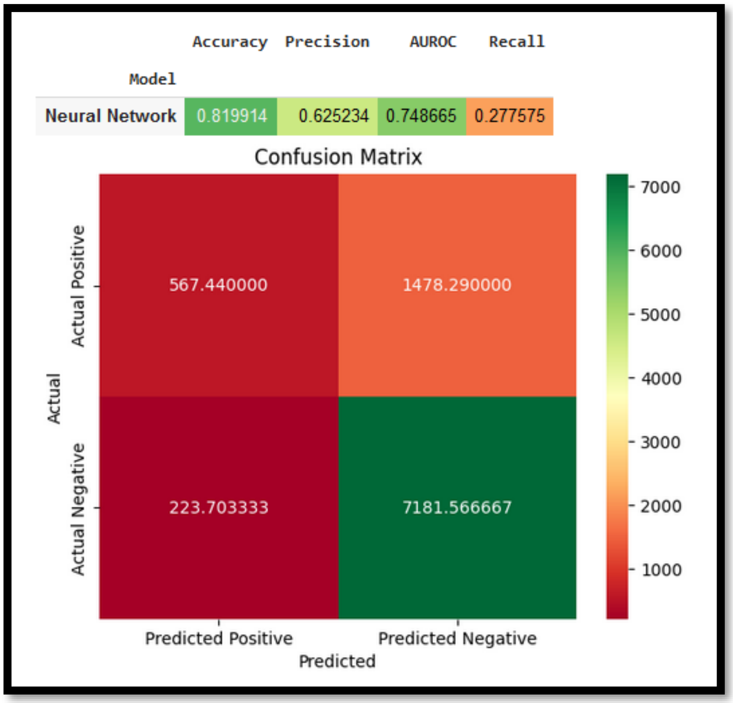
Model	Accuracy	Precision	AUROC	Recall
Gradient Boosting	0.824202	0.686399	0.754322	0.295228

The Gradient Boosting model achieves an accuracy of 0.8242, indicating that approximately 82.42% of loan default predictions are correct. The precision, which measures the proportion of correctly predicted defaults among all predicted defaults, is 0.6864. The AUROC score is 0.7543, indicating that the model has a good ability to distinguish between default and non-default instances. The recall is 0.2952, suggesting that the model captures around 29.52% of all actual defaults.

These results demonstrate that the Gradient Boosting model has the potential to identify a significant number of loan defaults correctly. However, it also produces a relatively high number of false positives, misclassifying some non-default instances as defaults. The model shows reasonable accuracy and precision, but the relatively low recall indicates that it may miss some actual defaults.

Neural Network

The Neural Network model correctly identifies 567.44 true positive cases, indicating its ability to accurately predict instances of default. However, it also produces a moderate number of false positive cases (223.70), misclassifying non-default instances. The model shows a high number of true negative cases (7181.57), correctly identifying non-default instances. However, it misclassifies 1478.29 instances as false negatives, indicating a relatively higher rate of missed default predictions compared to the Gradient Boosting model.



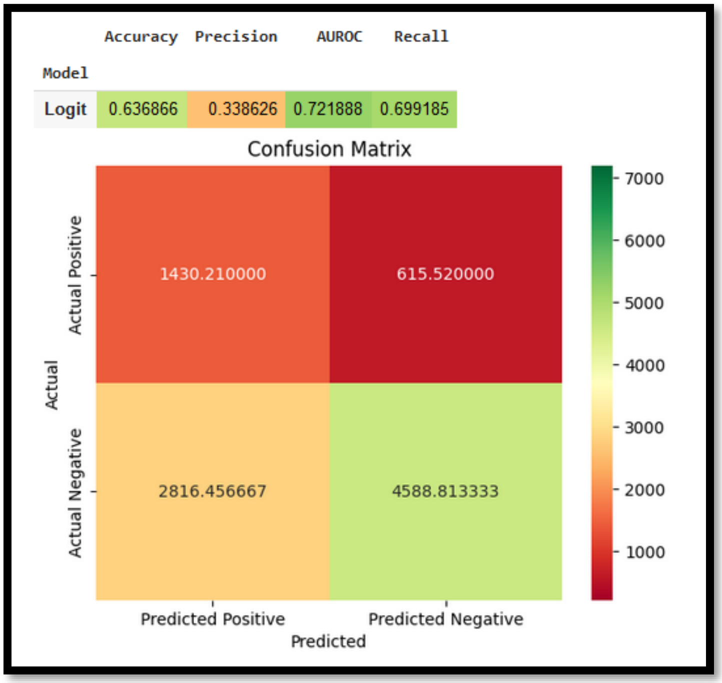
Model	Accuracy	Precision	AUROC	Recall
Neural Network	0.819914	0.625234	0.748665	0.277575

The model demonstrates an accuracy of 0.8199, suggesting that approximately 81.99% of loan default predictions are correct. The precision is 0.6252, indicating the proportion of correctly predicted defaults among all predicted defaults. The recall value is 0.2776, indicating that the model captures around 27.76% of all actual defaults. The AUROC score is 0.7487, indicating reasonable discriminatory performance.

These results indicate that the Neural Network model has strengths in correctly identifying loan defaults, but it also has limitations in controlling false positives and capturing all actual defaults. The model demonstrates a trade-off between correctly predicting defaults and minimizing false predictions.

Logit

The Logit model in this analysis correctly identifies 1430.21 true positive cases and produces 2816.46 false positive cases, indicating a relatively high rate of misclassification for non-default instances. It shows a high number of true negative cases (4588.81), correctly identifying non-default instances. However, it misclassifies 615.52 instances as false negatives, suggesting a higher rate of missed default predictions compared to the previous models. The model exhibits an AUROC value of 0.7219, indicating a moderately good ability to distinguish between default and non-default instances.



Model	Accuracy	Precision	AUROC	Recall
Logit	0.636866	0.338626	0.721888	0.699185

Confusion Matrix

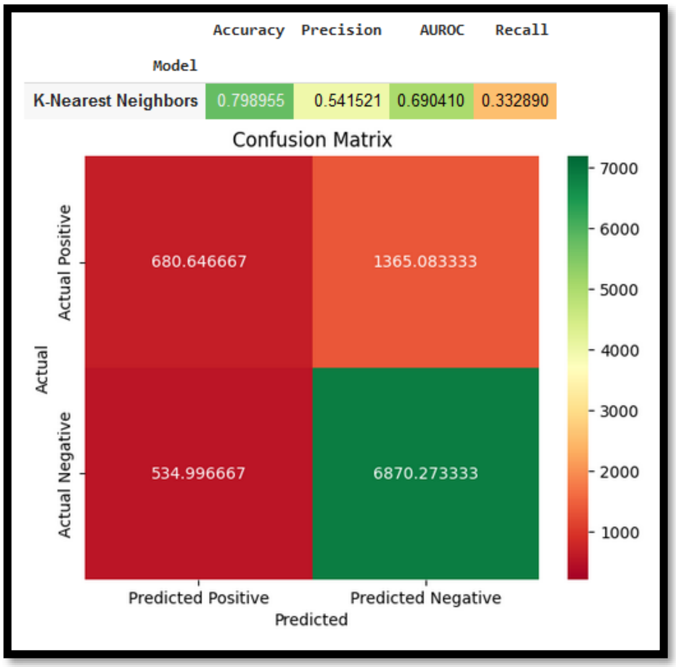
Actual \ Predicted	Predicted Positive	Predicted Negative
Actual Positive	1430.210000	615.520000
Actual Negative	2816.456667	4588.813333

The model demonstrates an accuracy of 0.6369, suggesting that approximately 63.69% of loan default predictions are correct. The precision is 0.3386, indicating the proportion of correctly predicted defaults among all predicted defaults. The recall value is 0.6992, indicating that the model captures around 69.92% of all actual defaults.

It's important to note that the Logit model's performance is comparatively poorer compared to the other models evaluated in this analysis. The relatively low accuracy, precision, and recall values suggest that the Logit model may not be as effective in predicting loan defaults compared to models like Gradient Boosting and Neural Network. However, the model still exhibits a moderate AUROC value, indicating some discriminatory power in distinguishing between default and non-default instances.

K-Nearest Neighbors

The K-Nearest Neighbors model demonstrates a relatively higher number of false positive cases (534.997) compared to other models, indicating a higher rate of misclassifying non-default instances. It correctly identifies a moderate number of true positive cases (680.647), but it also exhibits a relatively higher number of false negative cases (1365.083), suggesting a higher rate of missed default predictions. The model shows a high number of true negative cases (6870.273), correctly identifying non-default instances. The AUROC value for the K-Nearest Neighbors model is 0.6904, indicating its relatively lower ability to distinguish between default and non-default instances compared to the other models.



Model	Accuracy	Precision	AUROC	Recall
K-Nearest Neighbors	0.798955	0.541521	0.690410	0.332890

Confusion Matrix

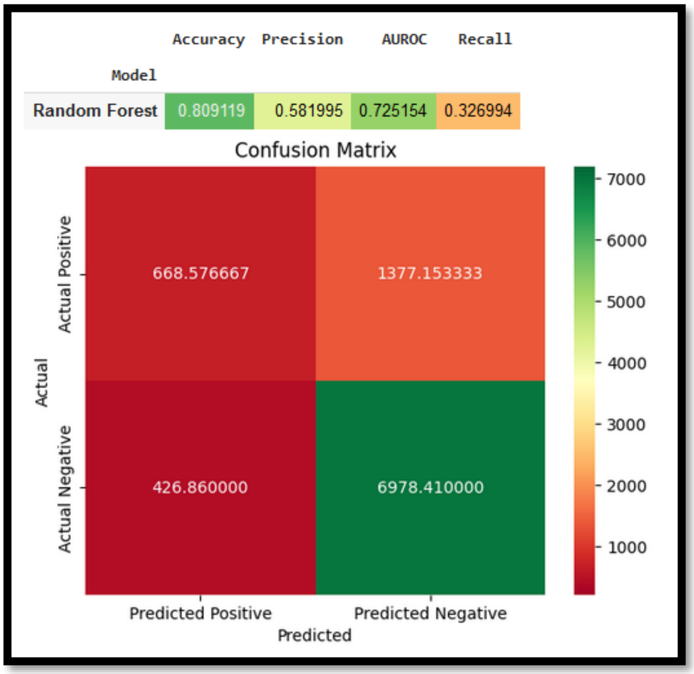
Actual \ Predicted	Predicted Positive	Predicted Negative
Actual Positive	680.646667	1365.083333
Actual Negative	534.996667	6870.273333

The model demonstrates an accuracy of 0.7990, suggesting that approximately 79.90% of loan default predictions are correct. The precision value is 0.5415 and the recall value is 0.3329, indicating that the model captures around 33.29% of all actual defaults.

Overall, the K-Nearest Neighbors model performs relatively lower in terms of distinguishing between default and non-default instances and predicting loan defaults accurately compared to the other models.

Random Forest

The Random Forest model exhibits a mixed performance in predicting loan defaults. It correctly identifies 668.58 instances of loan defaults. However, it also produces a relatively high number of false positives, misclassifying 426.86 instances as defaults when they are not. On the other hand, the model performs well in identifying non-default instances, as it correctly identifies 6,978.41 cases as non-defaults. However, it also misclassifies 1,377.15 instances as non-defaults when they are defaults.



Model	Accuracy	Precision	AUROC	Recall
Random Forest	0.809119	0.581995	0.725154	0.326994

In terms of overall accuracy, the Random Forest model achieves a score of 0.8091, indicating that approximately 80.91% of its loan default predictions are correct. The precision value of 0.5819 suggests that among all the predicted defaults, around 58.19% are truly defaults. The AUROC score of 0.7252 indicates that the model possesses reasonable discriminatory power in distinguishing between default and non-default cases. Lastly, the recall value of 0.3269 signifies that the model captures approximately 32.69% of all actual defaults.

Conclusion

In evaluating the performance of the models, accuracy serves as the main evaluation measure, providing an overall assessment of their predictive power. Additionally, analyzing the confusion matrix, AUROC, precision and recall helps to gain a deeper understanding of their performance.

Among the models considered, the Gradient Boosting model demonstrates the highest accuracy at 0.8242, indicating that it correctly predicts loan defaults approximately 82.36% of the time. This suggests that the model performs relatively well in identifying default instances and non-default instances. Examining the confusion matrix, it accurately classifies 8796 true negative cases, indicating its ability to identify non-default instances correctly. However, it misclassifies 1403 instances as false negatives, representing missed default predictions.

The Neural Network, Logit, K-Nearest Neighbors, and Random Forest models exhibit slightly lower accuracies, ranging from 0.63 to 0.81.

Considering AUROC, precision and recall the Gradient Boosting model continues to outperform the other models, indicating its superior ability to discriminate between default and non-default instances.

Web interface

This section outlines the setup process of the project's web interface, focusing on its primary attributes. The interface is developed using Voila in the Jupyter-Lab environment. Installation of Voila is the first step, which can be accomplished using the following command:

- ***pip** install voila*

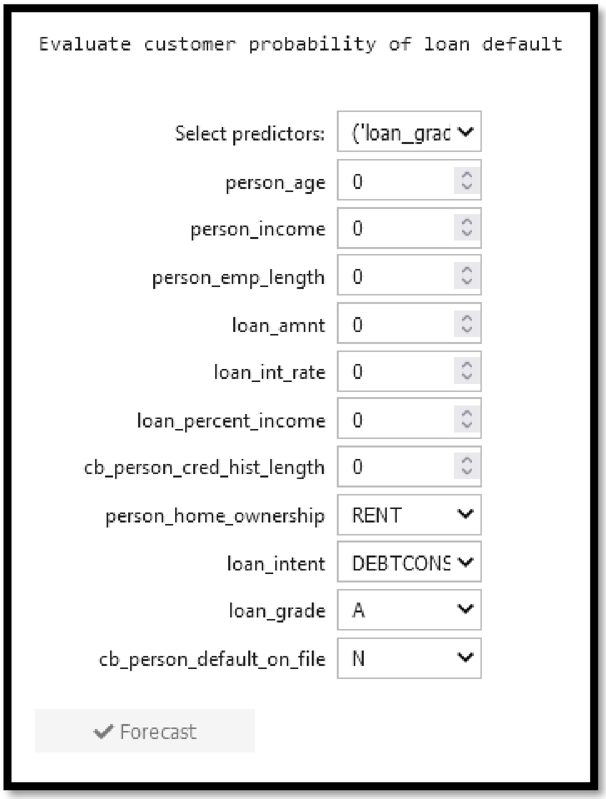
Post installation, the interface can be initiated in Jupyter-Lab. This requires the selection of 'View' from the toolbar, followed by 'Render Notebook with Voila'. This action prompts the '.ipynb' web app file to appear in an interactive Jupyter-Lab window.

Alternatively, the application can be run in a standard web browser. This requires execution of the following console command:

- ***Voila** [WebApp.ipynb]*

Execution of this command runs the app on a temporary server, facilitated by Voila, at localhost:8866.

The first feature to note is the 'Select predictors' option in the first cell. This allows selection from various predictor combinations. Once a combination has been chosen, corresponding parameters can be set. After adjusting all relevant parameters, a button can be clicked to display the probability of loan default for the selected combination. This probability is an average, computed from the probabilities provided by each individual model.



```
Evaluate customer probability of loan default
```

Select predictors:	('loan_grac
person_age	0
person_income	0
person_emp_length	0
loan_amnt	0
loan_int_rate	0
loan_percent_income	0
cb_person_cred_hist_length	0
person_home_ownership	RENT
loan_intent	DEBTCONS
loan_grade	A
cb_person_default_on_file	N

Adding to the previously mentioned information, it's critical to note that this application's successful operation depends on the models executed in the main project being saved. The absence of these saved models will hinder the proper functioning of the web interface. Hence, it's essential to run the project first, which will save the models as .pkl files. These files are subsequently accessed and utilized by the application.