Phase 4 Report: Feature Importance via Iterative Feature Removal

Cornell Stokes

May 1, 2025

1 Phase 4: Feature Importance Analysis

In Phase 4, the objective was to analyze the relative importance of the input features used in our machine learning model. This phase aimed to improve model performance and interpretability by identifying and potentially removing less informative features. The key methodology involved training individual models using one input feature at a time, followed by iteratively removing the least important features and comparing validation accuracy.

1.1 Step 1: Single-Feature Models

The first step was to train separate models using only one feature at a time. This helped determine which single features were most informative by observing the validation accuracy of each corresponding model.

Table 1: Single-Feature Model Performance

Feature	Validation Accuracy	Precision	Recall	F1 Score
person_age	77.88%	0.0	0.0	0.0
person_income	77.88%	0.0	0.0	0.0
person_emp_exp	77.88%	0.0	0.0	0.0
loan_amnt	77.88%	0.0	0.0	0.0
loan_int_rate	79.19%	61.48	15.87	0.252
loan_percent_income	**81.59%**	72.15	27.32	0.396
cb_person_cred_hist_length	77.88%	0.0	0.0	0.0
credit_score	77.88%	0.0	0.0	0.0
person_gender_male	77.88%	0.0	0.0	0.0
person_education_Bachelor	77.88%	0.0	0.0	0.0
person_education_Doctorate	77.88%	0.0	0.0	0.0
person_education_High School	77.88%	0.0	0.0	0.0
person_education_Master	77.88%	0.0	0.0	0.0
person_home_ownership_OTHER	77.88%	0.0	0.0	0.0
person_home_ownership_OWN	77.88%	0.0	0.0	0.0
person_home_ownership_RENT	77.88%	0.0	0.0	0.0
loan_intent_EDUCATION	77.88%	0.0	0.0	0.0
loan_intent_HOMEIMPROVEMENT	77.88%	0.0	0.0	0.0
loan_intent_MEDICAL	77.88%	0.0	0.0	0.0
loan_intent_PERSONAL	77.88%	0.0	0.0	0.0
loan_intent_VENTURE	77.88%	0.0	0.0	0.0
previous_loan_defaults_on_file_Yes	77.88%	0.0	0.0	0.0

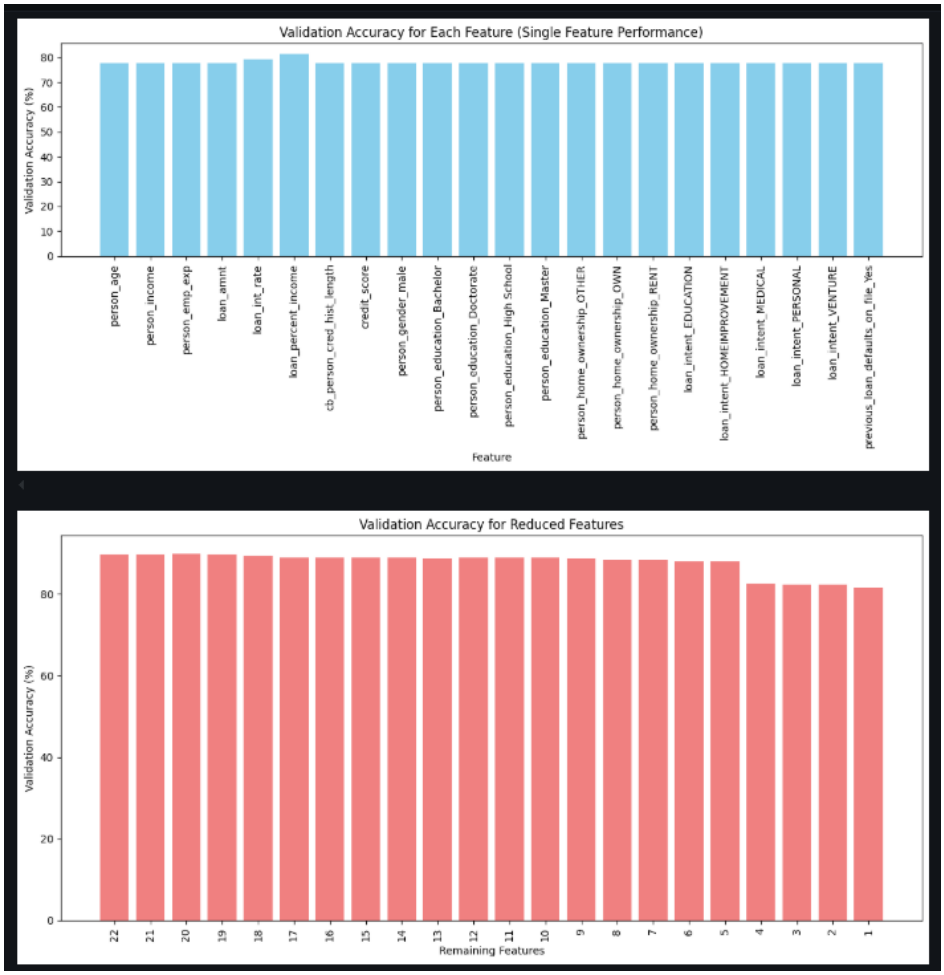
From the table above, we can see that the feature `loan_percent_income` resulted in the highest validation accuracy, indicating it is the most informative feature in isolation.

1.2 Step 2: Iterative Feature Reduction

In this step, models were trained with the progressively least important features removed. At each iteration, the validation accuracy and evaluation metrics were recorded to track performance changes.

Table 2: Reduced-Feature Model Performance

Remaining Features	**Validation Accuracy**	**Precision**	**Recall**	**F1 Score**
22	89.70%	77.71	74.94	0.763
21	89.67%	77.62	74.89	0.762
20	**89.77%**	77.98	74.89	0.764
19	89.70%	77.74	74.89	0.763
18	89.18%	75.93	74.79	0.754
17	88.90%	75.54	73.68	0.746
16	88.79%	75.47	73.08	0.743
15	88.79%	75.55	72.93	0.742
14	88.83%	75.57	73.18	0.744
13	88.73%	75.35	72.93	0.741
12	88.79%	75.08	73.83	0.744
11	88.78%	75.12	73.68	0.744
10	88.78%	75.24	73.43	0.743
9	88.58%	74.36	73.83	0.741
8	88.27%	75.26	69.96	0.725
7	88.23%	74.95	70.32	0.726
6	87.99%	74.76	69.01	0.718
5	87.96%	74.19	69.86	0.720
4	82.37%	68.40	37.72	0.486
3	82.24%	68.04	37.22	0.481
2	82.21%	68.29	36.56	0.476
1	81.59%	72.15	27.32	0.396



We observe that the best performance was achieved with 20 features, where the validation accuracy peaked at 89.77%. As features were progressively removed, accuracy and F1 score remained relatively stable until a sharp decline began around the 5-feature mark. This suggests that many features could be safely removed without degrading performance, and that feature reduction led to a more efficient and equally effective model.

1.3 Insights and Impact

This phase provided valuable insight into which features had the most influence on model predictions. The `loan_percent_income` and `loan_int_rate` features stood out as the most informative. In contrast, features like education, gender, and home ownership had minimal individual impact.

Removing non-informative features not only simplifies the model but can also lead to faster training times and better generalization. The feature-reduced models achieved similar or better accuracy than the original model with all features, supporting the effectiveness of this approach.

2 Conclusion

Through single-feature modeling and iterative feature removal, we identified the most important predictors and improved the model's efficiency without sacrificing performance. This phase underscores the value of feature selection as a key part of model tuning and provides a strong foundation for further model refinement and deployment.