Predicting Bank Loan Defaults DS4400 Final Project

Eduardo Ruiz-Garay, Samuel Baldwin

Project Goal

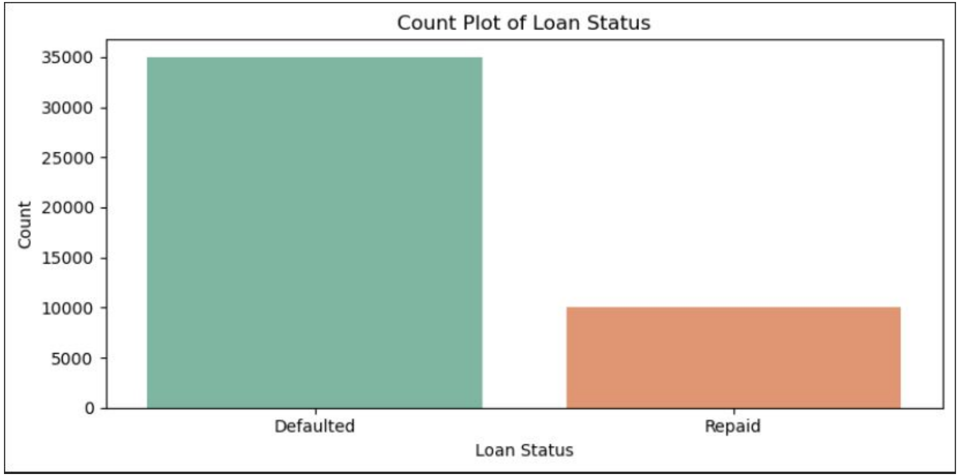
Predict loan defaults (can’t pay back) to reduce risk for institutions

Determine dataset validity and identify key features and distributions and hypothesis testing

Create and compare models and testing of models accuracy

Dataset

Kaggle dataset of 45,000 loan applicants with demographic and financial info

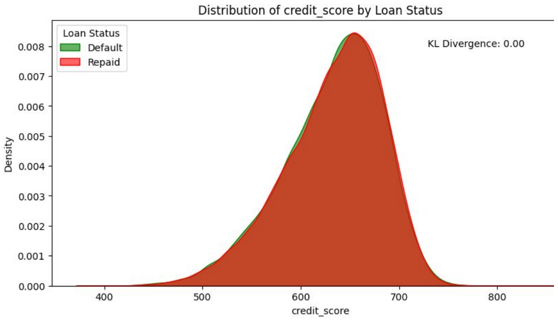


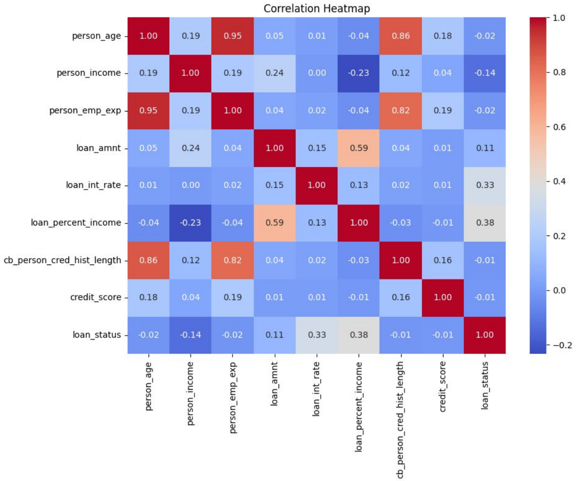
Classification predicting Loan Status
Features include (22 total):

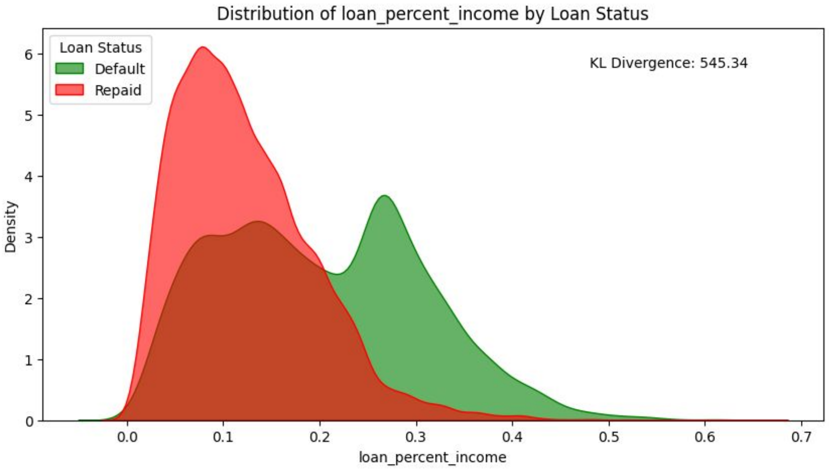
- Personal income
- Loan amount
- Credit History
- Loan History
- Home ownership status
- Education
- Age
- Gender

Data

- Use correlation matrix saw high correlation credit score, history w.r.t. status
- Used KDE different bandwidths to determine distributions and calculated overlap using simpsons and variance features high divergence loan percent income
- KL divergence both directions for symmetry
- Detecting multicollinearity and relationships
    - Good clean relationships and not too much skew except for 78% default and age >80







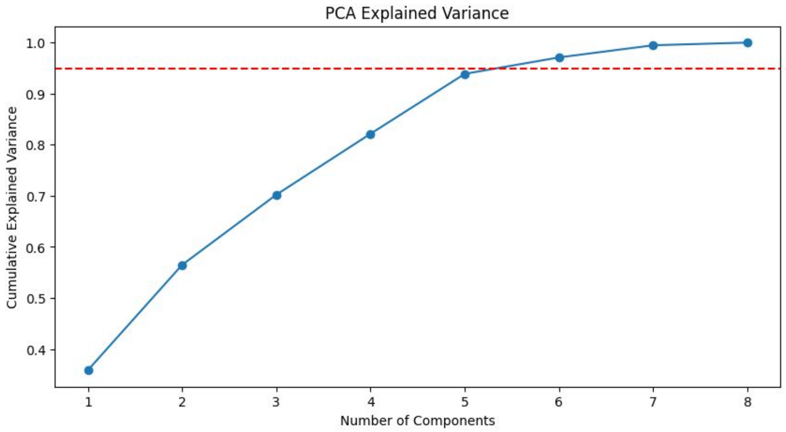
Preparation and Preprocessing

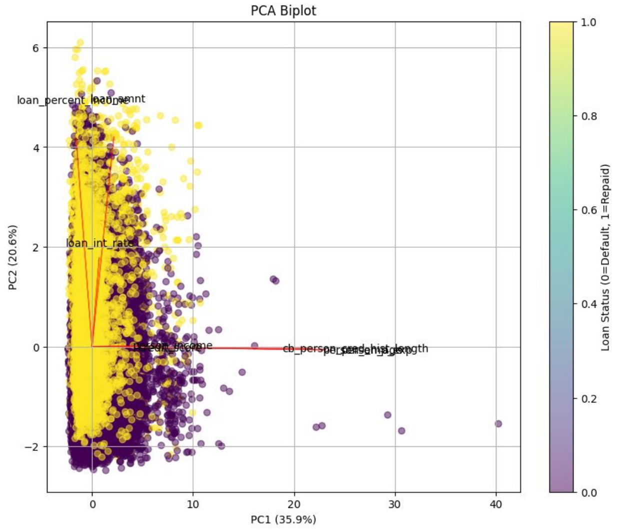
- Dropped any rows with null values and checked for duplicates
- Categorical features one hot encoded
- Numerical features
  - For Logistic Regression, KNN, Random Forest - min max scaled
  - For SVM, LDA, PCA - standard scaled
- Added intercept column for Logistic Regression
- Split observations into consistent train and test data for all models (80/20)

Models

PCA

- Calculated using covariance matrix and eigen decomposition
- High Overlap maintained in PCA projection data spread not clear due to high correlation were not clearly linearly separable in original features
- Determined to use LDA instead as used class labels and maximizes class seperation





Logistic Regression

- Implemented with gradient descent
- Learning rate: 0.1, Epochs: 1000
- 89% accuracy
- 68% recall for class 1 (repaid loan)
    - Predicting default when actually repaid
- Key features: previous defaults, loan to income ratio

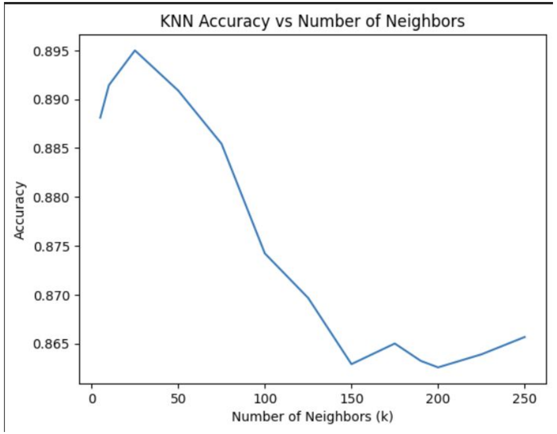
Classification Report:

	precision	recall	f1-score	support
0	0.91	0.95	0.93	6995
1	0.79	0.68	0.73	2005
accuracy			0.89	9000
macro avg	0.85	0.81	0.83	9000
weighted avg	0.88	0.89	0.88	9000

	Feature	Weight	Abs Weight
21	previous_loan_defaults_on_file	-3.256149	3.256149
22	intercept	-1.649544	1.649544
5	loan_percent_income	1.154655	1.154655
13	person_home_ownership_OTHER	0.978011	0.978011
4	loan_int_rate	0.920900	0.920900
15	person_home_ownership_RENT	0.860771	0.860771
10	person_education_Doctorate	0.820023	0.820023
17	loan_intent_HOMEIMPROVEMENT	0.683390	0.683390
18	loan_intent_MEDICAL	0.502514	0.502514
3	loan_amnt	-0.490486	0.490486
7	credit_score	-0.295191	0.295191
12	person_education_Master	0.274440	0.274440
19	loan_intent_PERSONAL	0.214913	0.214913

KNN

- Manually implemented using L2 norm to calculate distances
- Tested many k values, including square root of n (190)
    - Best performance at k=25 with 90% accuracy
- Low performance on repaid class (1)
- Improvement over logistic regression, but still lacking with minority class



Classification Report:

	precision	recall	f1-score	support
0	0.93	0.94	0.93	6995
1	0.78	0.74	0.76	2005
accuracy			0.90	9000
macro avg	0.85	0.84	0.85	9000
weighted avg	0.89	0.90	0.89	9000

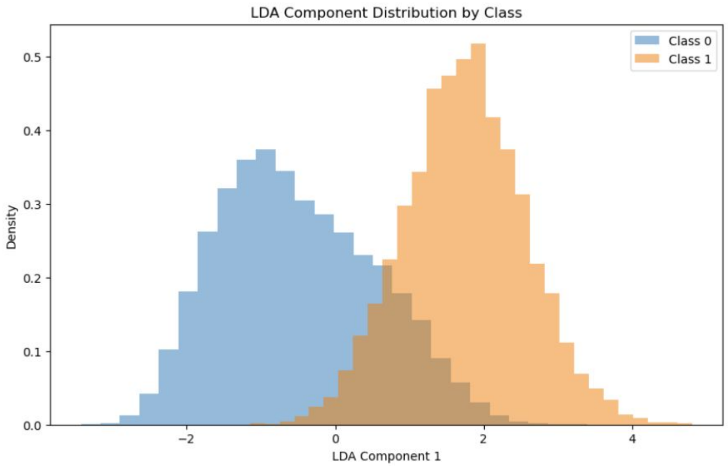
Random Forest

- Splits based on reducing entropy
- No max depth set
- 100 estimators (trees)
- 93% accuracy
- Best minority class prediction so far
- Most important features:
  - Previous loan defaults
  - Loan interest rate
  - Loan percent of income
  - Persons income

```
Classification Report:
              precision    recall  f1-score   support

           0       0.93      0.97      0.95      6995
           1       0.90      0.75      0.82      2005

    accuracy                           0.93      9000
   macro avg       0.91      0.86      0.89      9000
weighted avg       0.92      0.93      0.92      9000


Feature Importance:
                          Feature  Importance
21  previous_loan_defaults_on_file    0.284019
4                    loan_int_rate    0.140659
5              loan_percent_income    0.139325
1                    person_income    0.113094
3                        loan_amnt    0.059230
7                     credit_score    0.053026
15       person_home_ownership_RENT    0.047536
0                       person_age    0.031821
2                   person_emp_exp    0.028236
6       cb_person_cred_hist_length    0.026445
```

Linear Discriminant Analysis

- Reduces 22 dimensions to 1
- 1 component (hyperplane split) as binary classification requires this
- Normal distribution of classes with overlap
- 89% accuracy
- Struggles with minority class predictions (repaid)




```
Classification Report:
              precision    recall  f1-score   support

           0       0.93      0.93      0.93      6995
           1       0.76      0.75      0.75      2005

    accuracy                           0.89      9000
   macro avg       0.84      0.84      0.84      9000
weighted avg       0.89      0.89      0.89      9000
```

Support Vector Machine

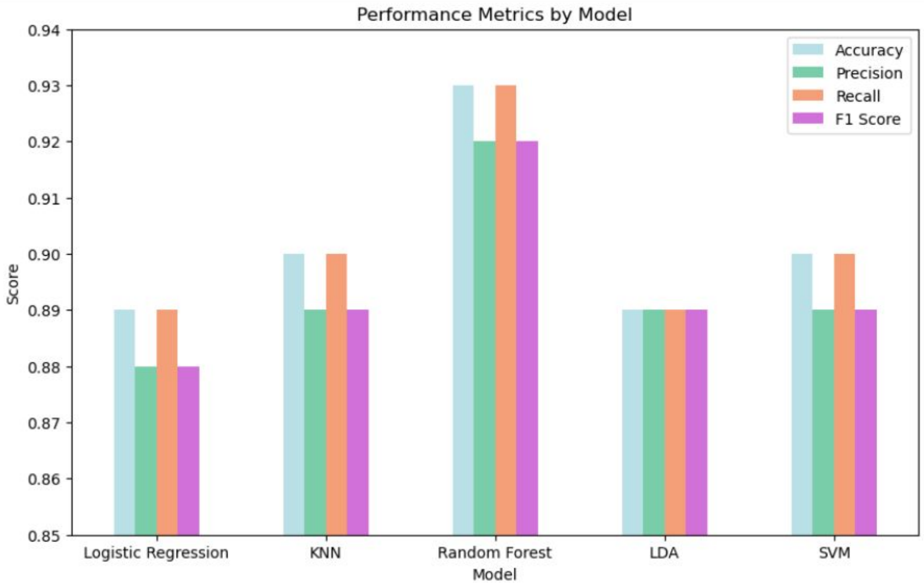
- Used soft margin (C=1) due to class overlap
- Hard margin SVM likely to overfit with overlapping classes
- 90% accuracy
- Weak with minority class

Classification Report:

	precision	recall	f1-score	support
0	0.93	0.94	0.93	6995
1	0.78	0.74	0.76	2005
accuracy			0.90	9000
macro avg	0.85	0.84	0.85	9000
weighted avg	0.89	0.90	0.89	9000

Results and Learnings

- The Random Forest was the best model in all measures
  - Specifically stood out in its ability to classify the minority class
  - Likely due to the robustness of RF and the randomness with the subsets of data used in each tree (bagging)
  - Random Forest is typically better that other models with correlated features
- The most important features were: previously defaulting on a loan, loan to income ratio, loan interest rate, income, and credit score
- Banks should be more cautious with making loans to people with worse metrics in these features, and anticipate lower repay rates



Future Work

- Fix class imbalance to improve predicting minority class
    - Undersampling the majority class or imputing synthetic minority class data into the training data could potentially fix this issue
- Feature engineering
    - Add and explore more features (disposable income estimate)
    - Test features interactions (credit * previous default)
- Test different encoding and scaling techniques
    - Ordinal encoding for education, maybe housing
- Optimize hyperparameters further
    - Gridsearch - test all combinations of hyperparameters
- Incorporate business cost
    - Are false negatives or false positives worse for profit?

Thank you!