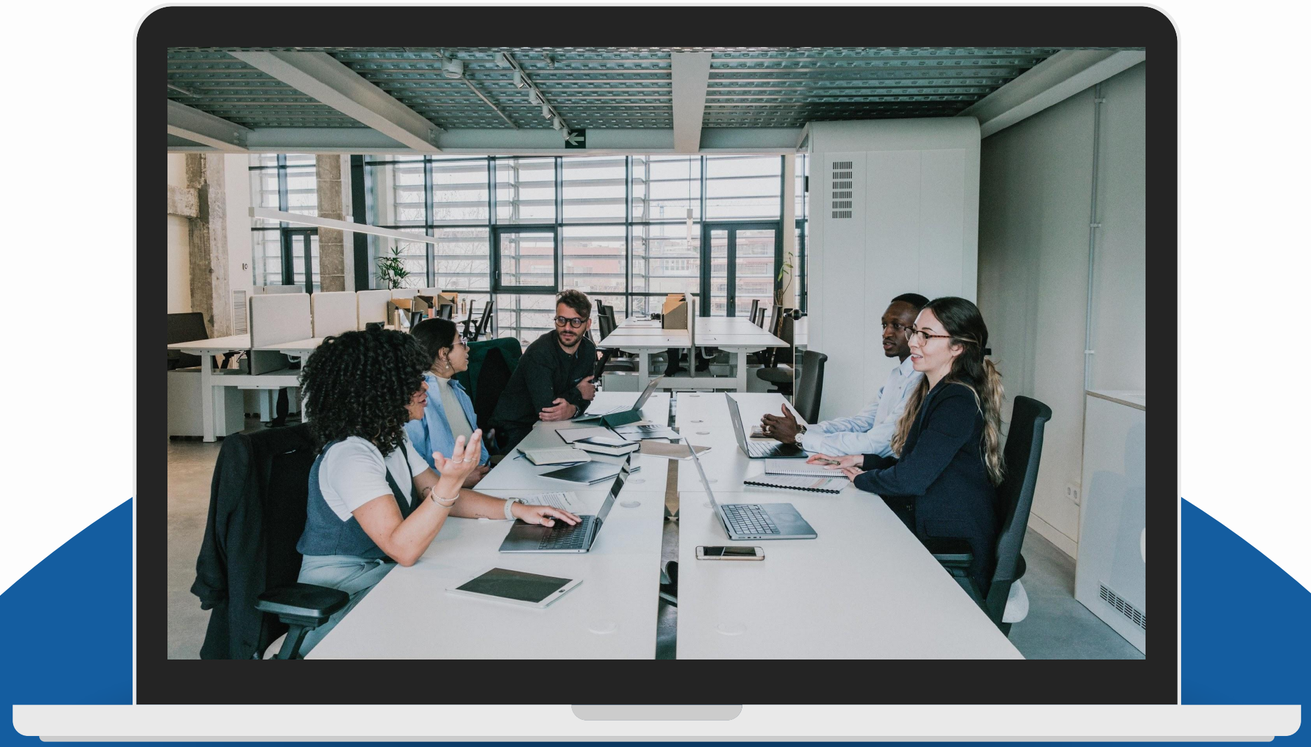
Loan Data Modeling of LendingClub

By: Uyemaa Gantulga, Ayush Meshram, Aakash Singh, and Melissa Yago



Overview

Introduction

- **LendingClub:** A financial services company that facilitates loan contracts.

- **Dataset (2013-2018)**
  Approved Loans Observations: 2.15 million

- **Goal:**
  To analyze LendingClub’s approved loan data to better understand factors that contribute to loan charge-offs.

Research Questions

1. **Features Impact:** Which combination of features from our initial EDA (interest rate, grade, annual income, etc.) provides the most reliable predictions of the 2013-2018 loan data?

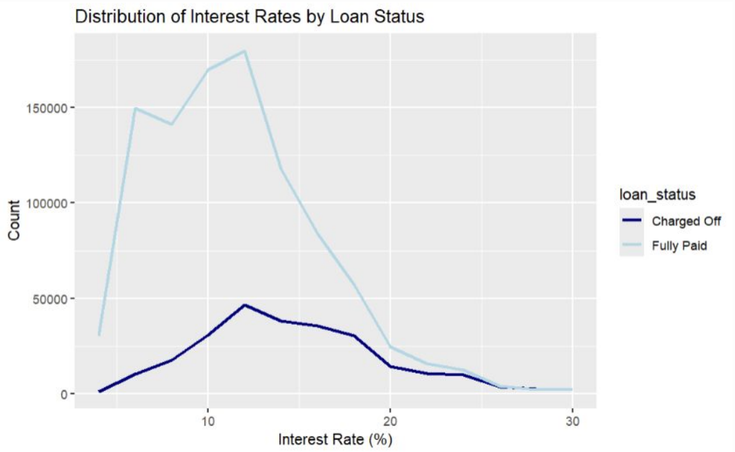
2. **Model Performance:** How do logistic regression and random forest models compare in their ability to predict loan charge-offs when trained on 2013-2015 data and tested on holdout sets from 2013-2015 and 2016-2018 data?

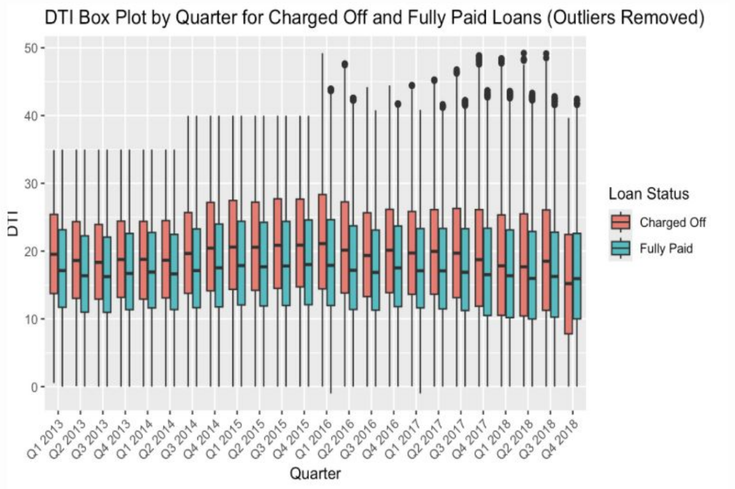
3. **Model Accuracy Across Timeframes:** How accurately can we predict loan charge-offs for loans issued between 2015-2018 that are still active and might charge-off in future, using our 2013-2018 trained models from Question #2?

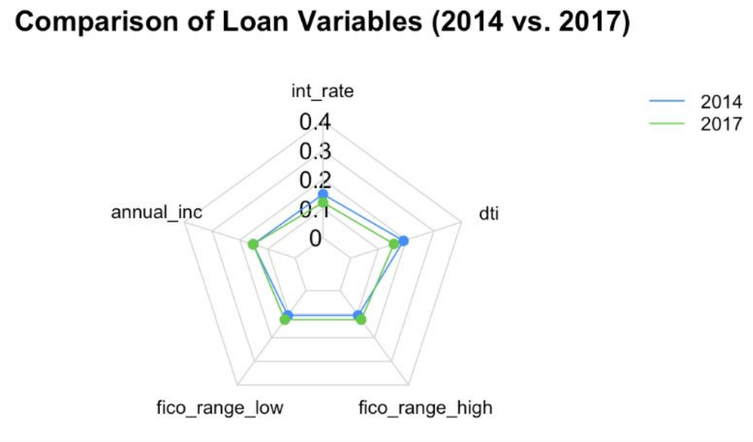
Question #1:
Features Impact

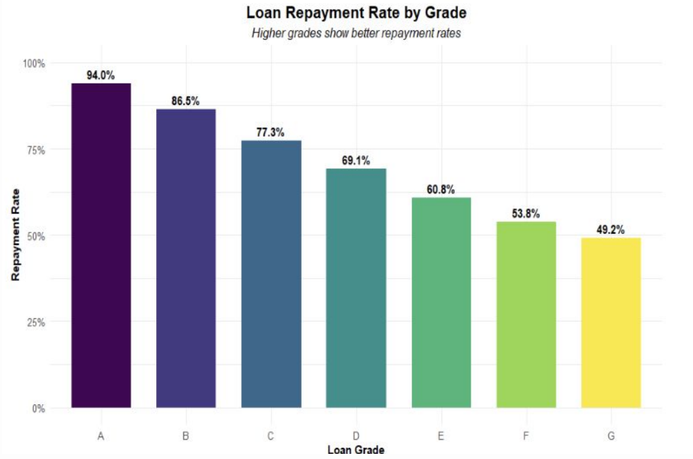
Which combination of features from our initial EDA (interest rate, grade, annual income, etc.) provides the most reliable predictions of the 2013-2018 loan data?

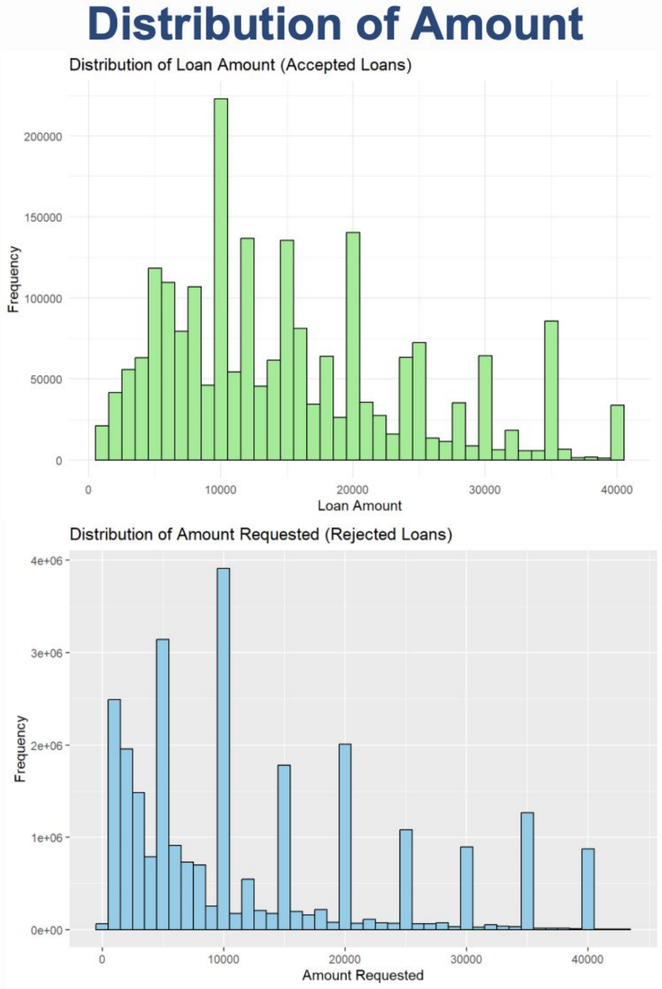
EDA Recap


Distribution of Interest Rates by Loan Status
Count
150000
100000
50000
0
10
20
30
Interest Rate (%)
loan_status
Charged Off
Fully Paid


DTI Box Plot by Quarter for Charged Off and Fully Paid Loans (Outliers Removed)
DTI
50
40
30
20
10
0
Q1 2013
Q2 2013
Q3 2013
Q4 2013
Q1 2014
Q2 2014
Q3 2014
Q4 2014
Q1 2015
Q2 2015
Q3 2015
Q4 2015
Q1 2016
Q2 2016
Q3 2016
Q4 2016
Q1 2017
Q2 2017
Q3 2017
Q4 2017
Q1 2018
Q2 2018
Q3 2018
Q4 2018
Quarter
Loan Status
Charged Off
Fully Paid


Comparison of Loan Variables (2014 vs. 2017)
int_rate
0.4
0.3
0.2
0.1
0
dti
fico_range_high
fico_range_low
annual_inc
2014
2017


Loan Repayment Rate by Grade
Higher grades show better repayment rates
Repayment Rate
100%
75%
50%
25%
0%
94.0%
86.5%
77.3%
69.1%
60.8%
53.8%
49.2%
A
B
C
D
E
F
G
Loan Grade


Distribution of Amount
Distribution of Loan Amount (Accepted Loans)
Frequency
200000
150000
100000
50000
0
0
10000
20000
30000
40000
Loan Amount
Distribution of Amount Requested (Rejected Loans)
Frequency
4e+06
3e+06
2e+06
1e+06
0e+00
0
10000
20000
30000
40000
Amount Requested

Final Logistic Regression & Random Forest Model Features

Independent Variables

- Loan Amount
- Interest Rate
- Grade
- Sub Grade
- DTI (Debt to Income Ratio)
- Employment Length
- Annual Income
- open_acc_6m
- Fico Scores
  - FICO Range High/Low
  - Latest FICO Range High/Low

Dependent Variable

- Loan Status

Question #2

Model Performance

How do different classification models (logistic regression and classification tree) compare in their ability to predict loan charge-offs when trained on 2013-2015 data and tested on holdout sets from 2013-2015 and 2016-2018 data?

Logistic Regression Model Performance

Initial Model Creation

- Did not include FICO predictors initially
- All predictors included in the model are statistically significant.
- **Coefficients**
  - **Loan Amount:** 0.13
  - **Interest Rate:** -0.62
  - **Grade G:** 5.29
  - **Employment Length<1 Year:** -0.24

Model Testing

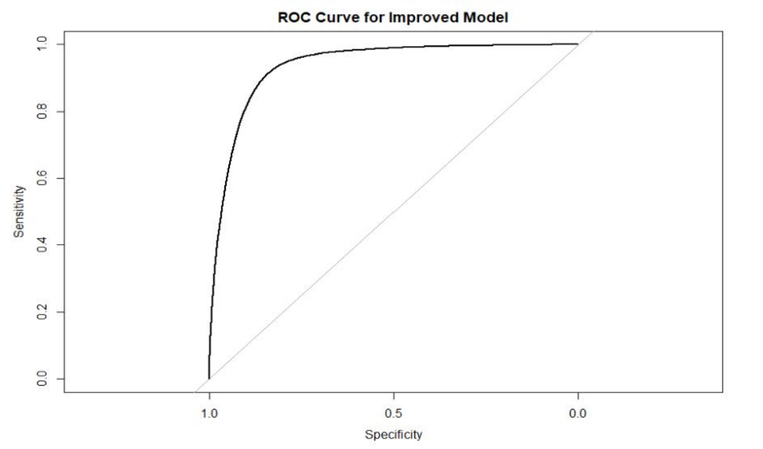
- **2013-2015 Test Data**
  - **Accuracy:** 81.3%
  - **Sensitivity:** 2.59%
  - **Specificity:** 99.40%
  - **AUC:** 0.705
- **2016-2018 Test Data**
  - **Accuracy:** 77.4%
  - **Sensitivity:** 1.01%
  - **Specificity:** 99.69%
  - **AUC:** 0.696

Model Accuracy Improvement

- **Training Data**
  - Combined Test datasets:
    - (2013-2015)
    - (2016-2018)
- **Final Selected Features**
  - loan_amnt, int_rate, grade, sub_grade, emp_length, annual_inc, dti, FICO ranges, open_acc_6m, loan_status
- **Data Processing**
  - MinMax normalization
  - 70/15/15 train-eval-test split
  - Feature engineering: 55 new features
- **Missing Data Treatment**
  - Median imputation (numeric)
  - Mode imputation (categorical)
  - No missing loan_status found

Logistic Regression Model Performance

Hyperparameter Tuning



- **What We Tuned**
  - We tested 15 different model settings by adjusting two key parameters:
    - Alpha (mixing ratio): Controls the blend between Ridge and Lasso
    - Lambda (strength): Controls how strongly the model penalizes complex patterns

- **Testing Approach**
  - Used 3 alpha values: 0 (Ridge), 0.5 (Mix), 1 (Lasso)
  - Tested 5 lambda values from 0.001 to 0.1
  - Used 5-fold cross-validation for reliable results
  - Selected best model based on ROC score

Model Performance Metrics

Accuracy
85.70%
CI: 85.50%-85.90%

F1 Score
90.60%

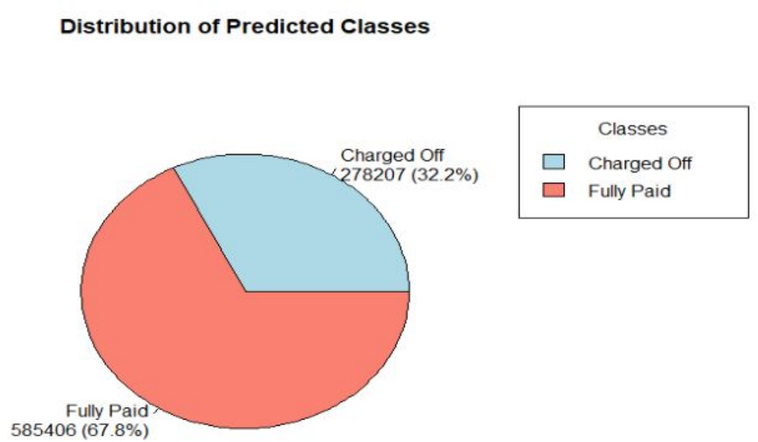
Specificity
90.90%

AUC
93.60%

Sensitivity
84.50%

Precision
97.60%

Predicting Loan Outcomes



Distribution Overview

Fully Paid (Expected): 585,406 loans (67.8%)

Charged Off (Expected): 278,207 loans (32.2%)

Total Active Loans: 863,613

Key Insights

Majority of loans (2/3) predicted to be paid fully

About 1/3 of loans flagged as potential charge-offs

Prediction based on optimal threshold from model

Random Forest

Why Random Forest

- Handling Non-Linear Relationships
- Feature Importance
- Robustness to Data Characteristics
- Overfitting and Regularization
- Class Imbalance

Handling Missing Values

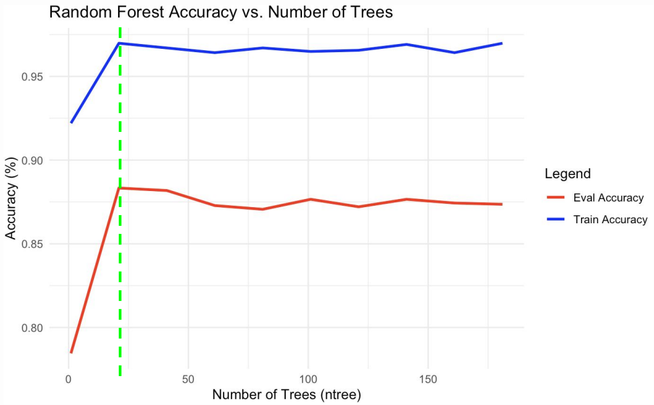
- KNN Imputation on Numerical variables
- Mode Imputation on Categorical Variables
- Removed observations which had missing loan_status

Data Preparation

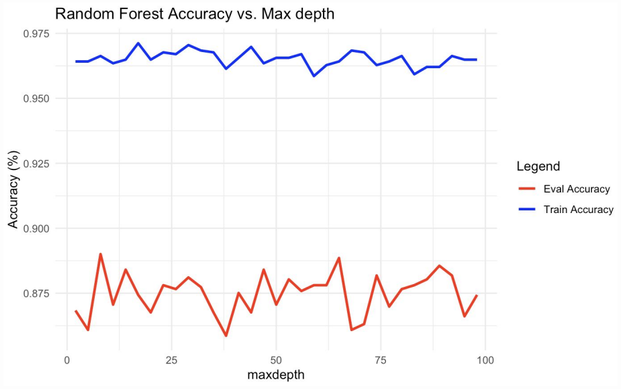
- Normalised the data using MinMax
- 70% train data, 15% evaluation data and 15% Test data

Hyper Parameter Tuning

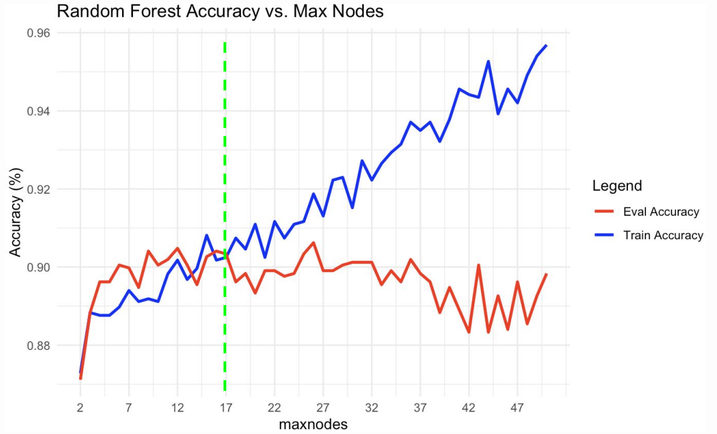
- ntree : 25



- max_depth : default



- Max_nodes : 17



- Prob_threshold: 0.7

Threshold <dbl>	Precision <dbl>	Recall <dbl>	F1 <dbl>
0.0	0.2033213	1.00000000	0.3379336
0.1	0.6982398	0.86587325	0.7730735
0.2	0.7005778	0.86200269	0.7729521
0.3	0.7048330	0.85823745	0.7740074
0.4	0.7093961	0.85460386	0.7752592
0.5	0.7184072	0.84317649	0.7758074
0.6	0.7357871	0.81955818	0.7754167
0.7	0.7526005	0.79441270	0.7729415
0.8	0.7708723	0.73711788	0.7536173
0.9	0.8080633	0.38155296	0.5183503

Random Forest

Metrics

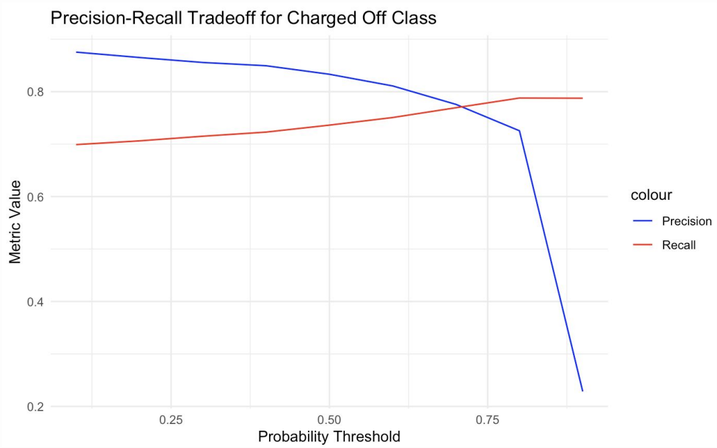
Class	Metrics	Train	Eval
Combined	Accuracy	0.9044	0.9054
Charged Off	Precision	0.833	0.835
Fully Paid		0.922	0.923
Charged Off	Recall	0.733	0.735
Fully Paid		0.955	0.956
Charged Off	F1	0.7801	0.782
Fully Paid		0.938	0.939

- Huge gap between Precision and Recall of Charged Off Predictions

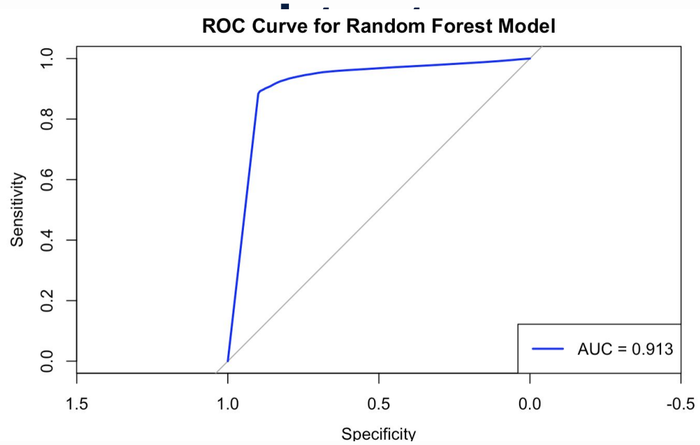
```
$eval_confusion_matrix
             Actual
Predicted     Charged Off Fully Paid
  Charged Off       31143      10609
  Fully Paid         6836     138204
```

Focus on Recall

- Recall measures the ability to correctly identify all actual charge-offs (true positives).
- Missing a charge-off (false negative) means the system fails to flag a potentially high-risk loan, which could lead to financial losses for the lender.
- High recall ensures that most risky loans are identified, reducing the likelihood of unpredicted financial defaults.
- We are choosing Probability Threshold as 0.7



Testing the Model on Test




```
             Precision    Recall
Charged Off 0.7731904 0.7645941
Fully Paid  0.9392463 0.9419491
```

Question #3:
Model Prediction Across Timeframes

How accurately can we predict loan charge-offs for Lending Club loans issued between 2015-2018 that are still active and might charge-off in future, using our 2013-2018 trained models from Question #2?

Model Predictions On 2015-2018 Active Loans

Logistic Regression

Outcome	Probability Threshold 0.5	Probability Threshold 0.7
Charged Off	115,242 (~13%)	76,601 (~8%)
Fully Paid	748,371	787,012

Random Forest

Outcome	Probability Threshold 0.5	Probability Threshold 0.7
Charged Off	196,234 (~22%)	164,248 (~19%)
Fully Paid	667,379	699,365

Conclusion

Key insights:

- Effective prediction relies on having FICO scores, loan grades and interest rates into modeling strategies
- Income levels and loan amounts are secondary predictors
- Continuous model refinement is necessary for evolving economic and borrower trends

Recommendations:

- Adopting hybrid modeling approach like combining logistic regression and random forest to maximize interpretability and recall
- Regularly update models with new data and external factors to maintain accuracy

01

Interest rates and FICO scores are the strongest predictors of loan charge offs, while loan amount does not significantly affect loan status.

02

Model performance comparison:
Logistic regression offers interpretability and stable accuracy. Random forest is excellent in recall, effectively identifying high-risk loans, but precision-recall trade-offs must be carefully managed.

03

Prediction accuracy across time frame:
Models trained on 2013-2015 data show low accuracy when applied to 2016-2018 loans due to borrower behaviour and policy changes

Post 2015 lending practices significantly impacted loan risk profiles, favoring lower-risk applicants.

THANK YOU!

Questions??