Business Case:
Prediction of Charged Off Loans

Applicant: Pedro Martinez

The Problem and Solution Proposal

- Objective: design a model that predicts Charged Off Loans using LC’s database
- Solution Structure
  - Exploratory Data Analysis
  - Analysis and data cleaning (missing values, “o.h.e.” on categorical variables)
  - Feature Selection (Forward Selection, Gradient Boosted Weights and Regularization)
  - Model Training
    - three models: Logistic Regression, Gradient Boosting, K-Nearest Neighbors
    - GridSearch for HyperParameters and Cross Validation for overfitting analysis
    - Performance analysis: Confusion Matrices and ROC Curves
  - Model Testing
    - Performance analysis: Confusion Matrices and ROC Curve

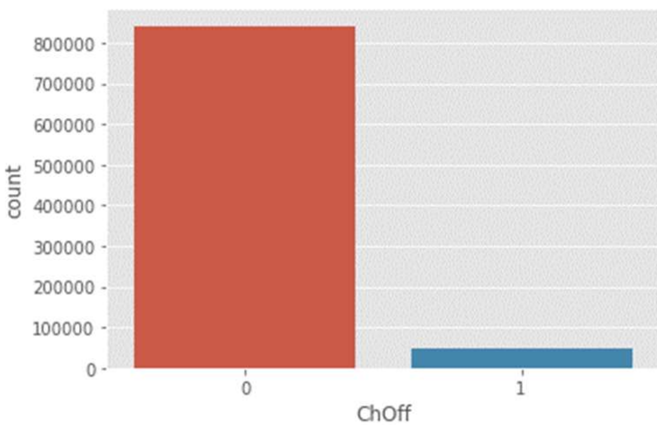
Dataset

- Train set: 887379 observations by 74 variables
- Test set: 759338 observations by 72 variables
- Columns with more than half of observations with missing values:

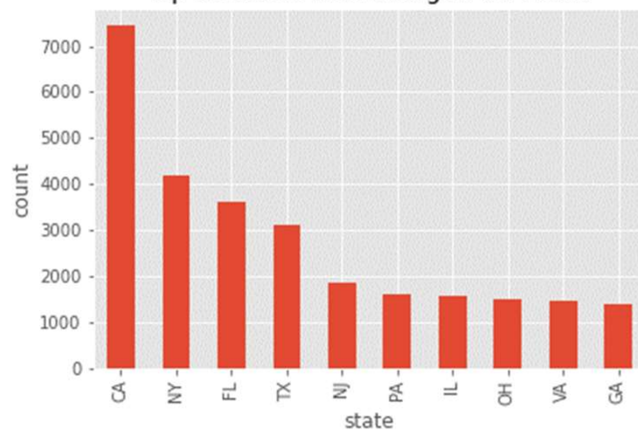
desc	761351
mths_since_last_delinq	454312
mths_since_last_record	750326
mths_since_last_major_derog	665676
annual_inc_joint	886868
dti_joint	886870
verification_status_joint	886868
open_acc_6m	866007
open_il_6m	866007
open_il_12m	866007
open_il_24m	866007
mths_since_rcnt_il	866569
total_bal_il	866007
il_util	868762
open_rv_12m	866007
open_rv_24m	866007
max_bal_bc	866007
all_util	866007
inq_fi	866007
total_cu_tl	866007
inq_last_12m	866007

Exploratory Data Analysis

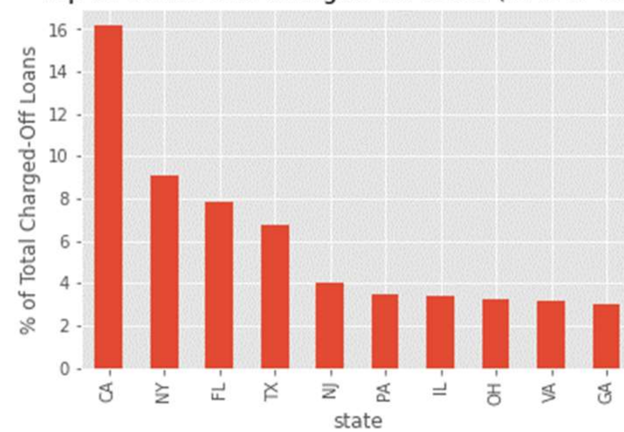
Number of Non Charged-Off vs Charged-Off Loans



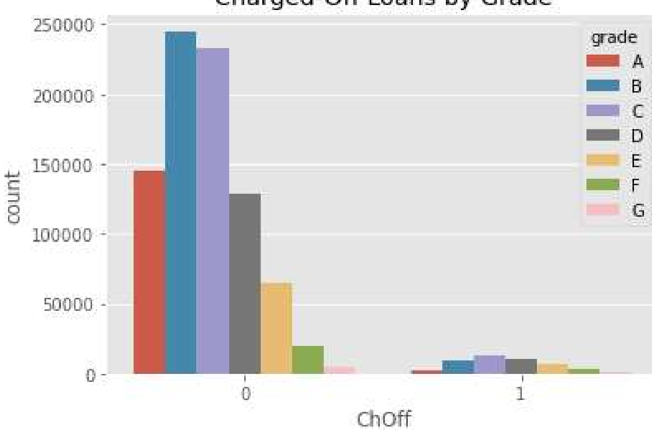
Top 10 states with Charged-Off Loans



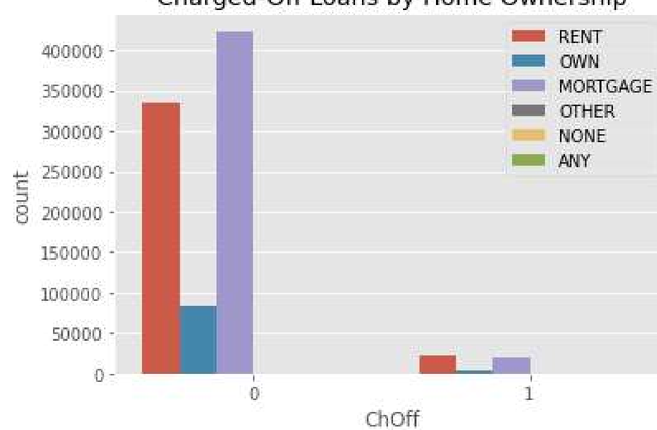
Top 10 states with Charged-Off Loans (in % of Totals)



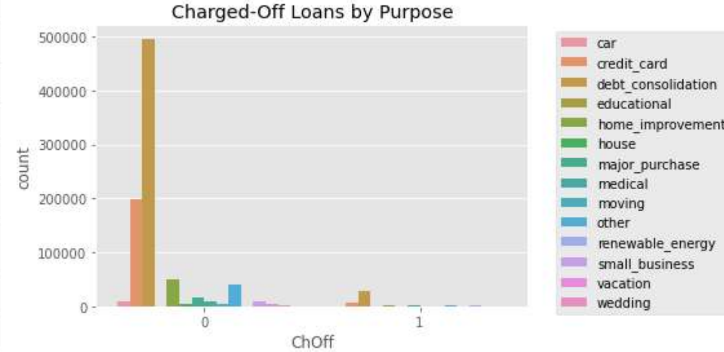
Charged-Off Loans by Grade



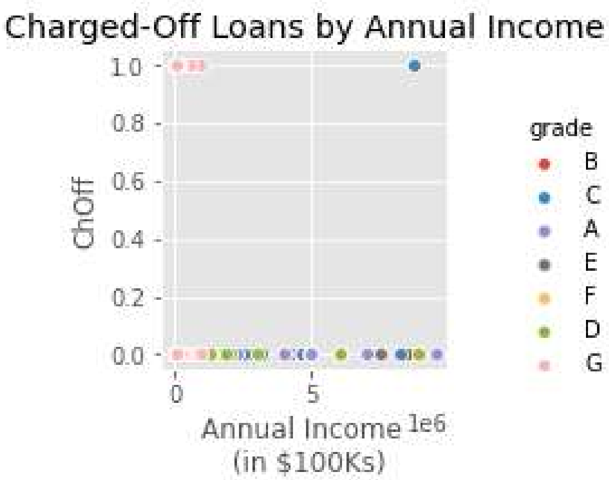
Charged-Off Loans by Home Ownership

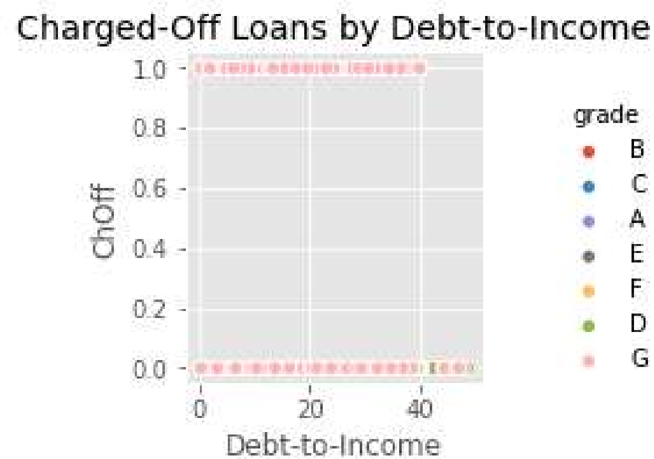


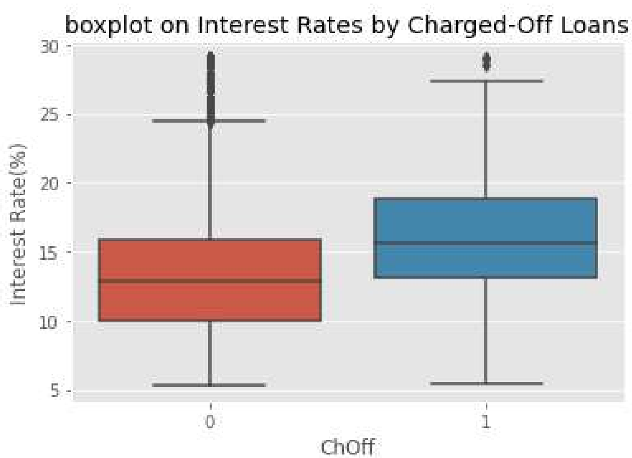
Charged-Off Loans by Purpose

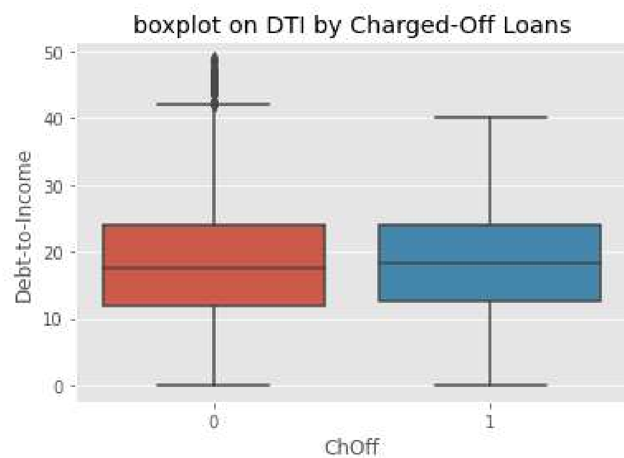


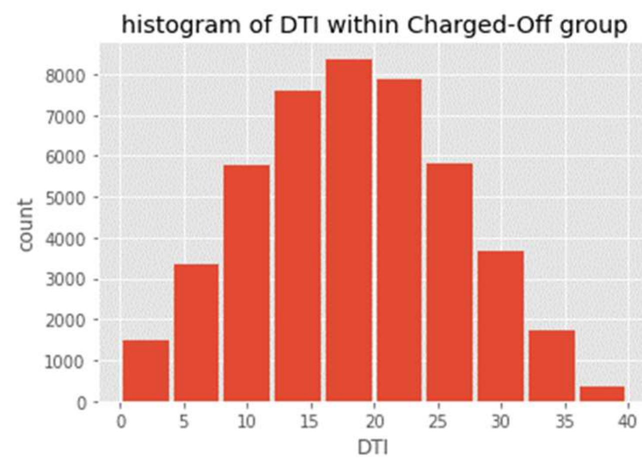
Exploratory Data Analysis (2)











Exploratory Data Analysis (conclusion)

- Average Loan Rate for Charged-Off borrowers may show common higher cost:

Avg Loan Rate per Home Ownership
(for Charged-Off and none's)

ChOff	0	1
home_ownership		
ANY	14.2	NaN
MORTGAGE	12.8	15.9
NONE	14.3	15.3
OTHER	13.2	14.3
OWN	13.1	16.1
RENT	13.4	16.1

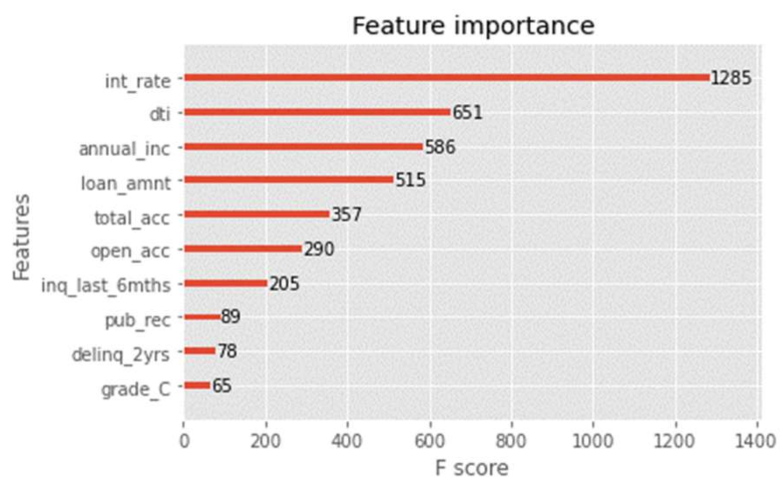
- In summary, a preliminary analysis shows no noticeable common characteristics among Charged-Off Loan borrowers compared to others.

Feature Selection

Several selection techniques were performed

- Regularization (for Logistic Regression): loan amount, interest rate, annual income, Debe-to-Income, delinquencies in the last 2 years, inquiries last 6 months, employment length, number of derogatory public records, number of open credit lines in the borrower's credit file.

- Gradient Boosted Weights (for Gradient Boosting):



Logistic Regression

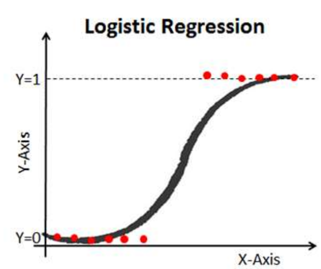
- Called logistic since it uses the Logit function $log_e\left(\frac{p}{1-p}\right)$
- The regression is in the form of

$$p = \left(\frac{1}{1 + e^{-\beta \cdot X}}\right)$$

where β is the coefficients matrix representing the log-odds for p =1

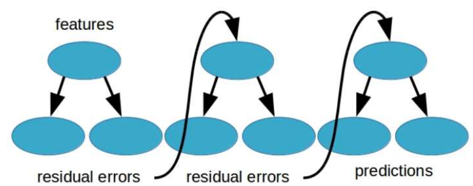
X is the feature or explanatory variables matrix

p is the probability of the target variable being 1

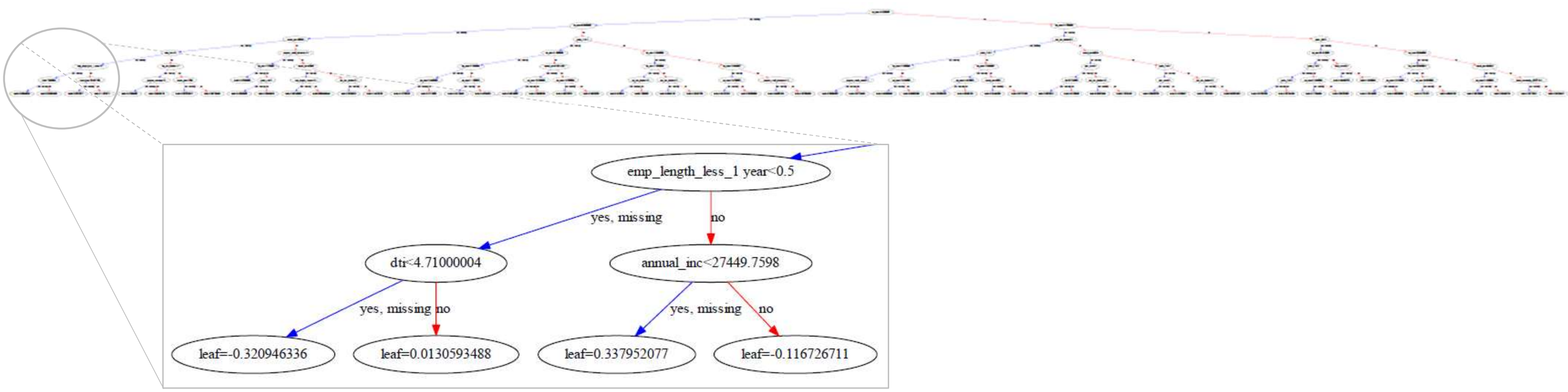


Gradient Boosting

A gradient boosting tress is an ensemble learning technique which predicts in the form of an ensemble of decision trees where the results of the each base-learner are combined to generate the final estimate.

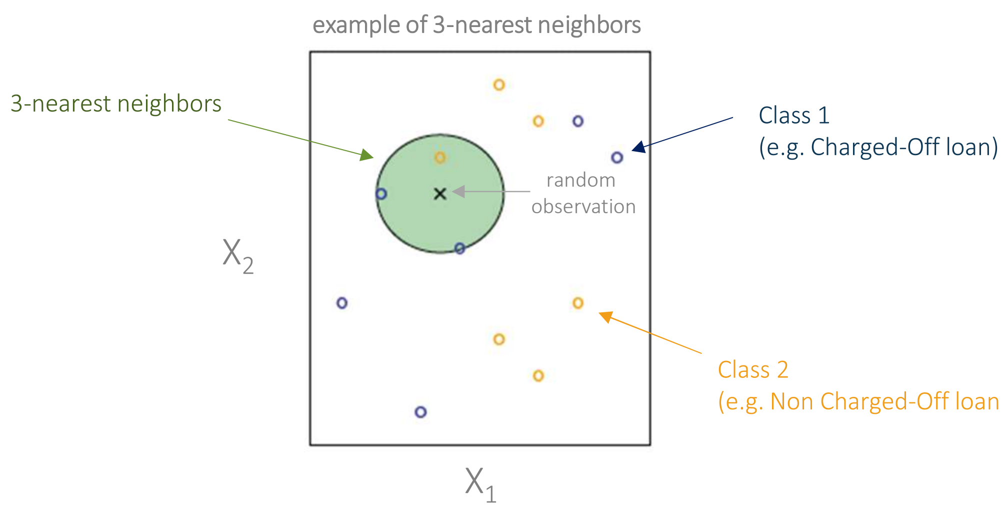


XGBoost Plot from Charged-Off Model:

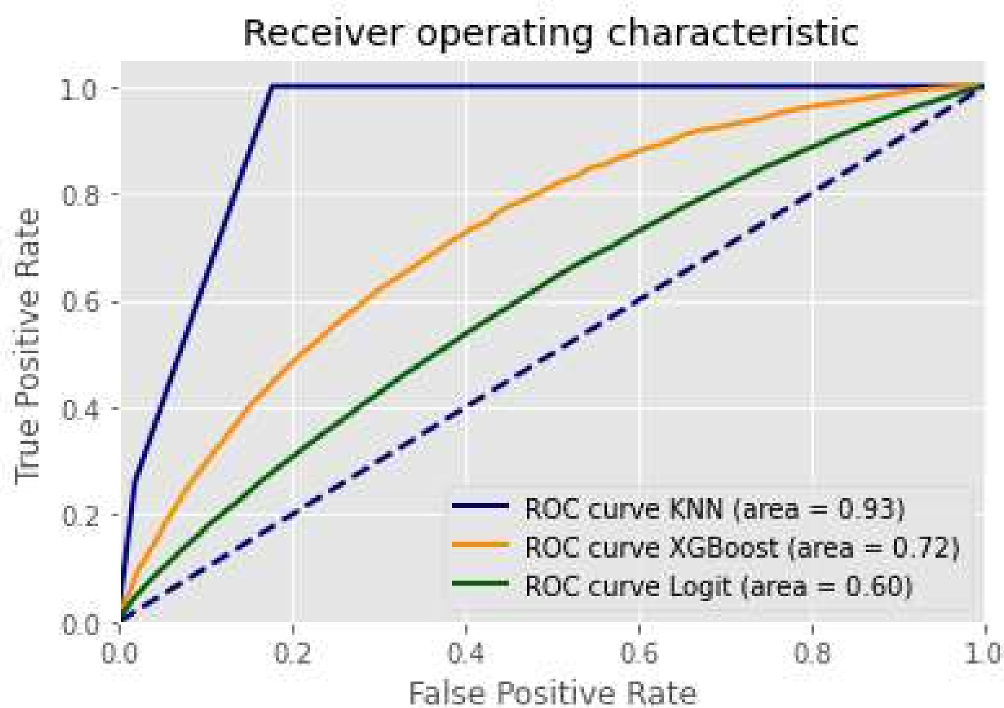


K-Nearest Neighbor

Given a defined 'number of neighbors' (K), for every pair combination of features (e.g. X_1 and X_2), this model analyzes for each random observation the K-nearest points and identifies their class to then estimate the very class of it based on the distribution of their neighbors (i.e. conditional probability).



Model Training and Performance Comparison



Confusion Matrix

	Predicted Negative	Predicted Positive
Actual Negative	TRUE NEGATIVE	FALSE POSITIVE
Actual Positive	FALSE NEGATIVE	TRUE POSITIVE

	Predicted Negative	Predicted Positive
Actual Negative	$1 - \alpha$	α
Actual Positive	β	$1 - \beta$

Results – Train set

Regularized Logistic Regression	Predicted *No Charged-Off*	Predicted *Charged-Off*
Actual *No Charged-Off*	841,201	143
Actual *Charged-Off*	45,935	71

Gradian Boosting	Predicted *No Charged-Off*	Predicted *Charged-Off*
Actual *No Charged-Off*	840,904	440
Actual *Charged-Off*	45,773	233

K-Nearest Neighbors	Predicted *No Charged-Off*	Predicted *Charged-Off*
Actual *No Charged-Off*	840,323	1,021
Actual *Charged-Off*	44,218	1,788

Results – Test set

Regularized Logistic Regression	Predicted *No Charged-Off*	Predicted *Charged-Off*
Actual *No Charged-Off*	715,350	14
Actual *Charged-Off*	43,618	0

Gradian Boosting	Predicted *No Charged-Off*	Predicted *Charged-Off*
Actual *No Charged-Off*	715,054	310
Actual *Charged-Off*	43,587	31

K-Nearest Neighbors	Predicted *No Charged-Off*	Predicted *Charged-Off*
Actual *No Charged-Off*	713,791	1,573
Actual *Charged-Off*	43,520	98

Classification Report

- Precision: $\frac{\text{TP}}{(\text{TP} + \text{FP})}$

- Recall: $\frac{\text{TP}}{(\text{TP} + \text{FN})}$

- F1score: $2 \cdot \frac{\text{precision} \cdot \text{recall}}{(\text{precision} + \text{recall})}$

- High precision: Predicted most of *Non-Charged Off Loans* correctly

- High recall: Not many actual *Charged Off Loans* predicted as *Non-Charged Off*

Classification Report: Train and Test sets

TRAIN

REGULARIZED LOGISTIC REGRESSION

	precision	recall	f1-score	support
No Charged-Off	0.95	1.00	0.97	841,344
Charged-Off	0.33	0.00	0.00	46,006
accuracy			0.9481	887,350
macro avg	0.64	0.50	0.49	887,350
weighted avg	0.92	0.95	0.92	887,350

GRADIENT BOOSTING

	precision	recall	f1-score	support
No Charged-Off	0.95	1.00	0.97	841,344
Charged-Off	0.35	0.01	0.01	46,006
accuracy			0.9479	887,350
macro avg	0.65	0.50	0.49	887,350
weighted avg	0.92	0.95	0.92	887,350

KNN

	precision	recall	f1-score	support
No Charged-Off	0.95	1.00	0.97	841,344
Charged-Off	0.64	0.04	0.07	46,006
accuracy			0.9490	887,350
macro avg	0.79	0.52	0.52	887,350
weighted avg	0.93	0.95	0.93	887,350

TEST

REGULARIZED LOGISTIC REGRESSION

	precision	recall	f1-score	support
No Charged-Off	0.94	1.00	0.97	715,364
Charged-Off	0.00	0.00	0.00	43,618
accuracy			0.9425	758,982
macro avg	0.47	0.50	0.49	758,982
weighted avg	0.89	0.94	0.91	758,982

GRADIENT BOOSTING

	precision	recall	f1-score	support
No Charged-Off	0.94	1.00	0.97	715,364
Charged-Off	0.09	0.00	0.00	43,618
accuracy			0.9422	758,982
macro avg	0.52	0.50	0.49	758,982
weighted avg	0.89	0.94	0.91	758,982

KNN

	precision	recall	f1-score	support
No Charged-Off	0.94	1.00	0.97	715,364
Charged-Off	0.06	0.00	0.00	43,618
accuracy			0.9406	758,982
macro avg	0.50	0.50	0.49	758,982
weighted avg	0.89	0.94	0.91	758,982

conclusion 1: in terms of Accuracy, KNN model performs better in the Train set

conclusion 2: in terms of Accuracy, Regularized Logistic Regression model performs better in the Test set

Further Improvement

- Feature Engineering:
  - Standardization
  - Normalization
  - log-changes
- Outlier Analysis
- Other Classifying Models

Lending Club

Business Case:

Prediction of Charged Off Loans

Applicant: Pedro Martinez