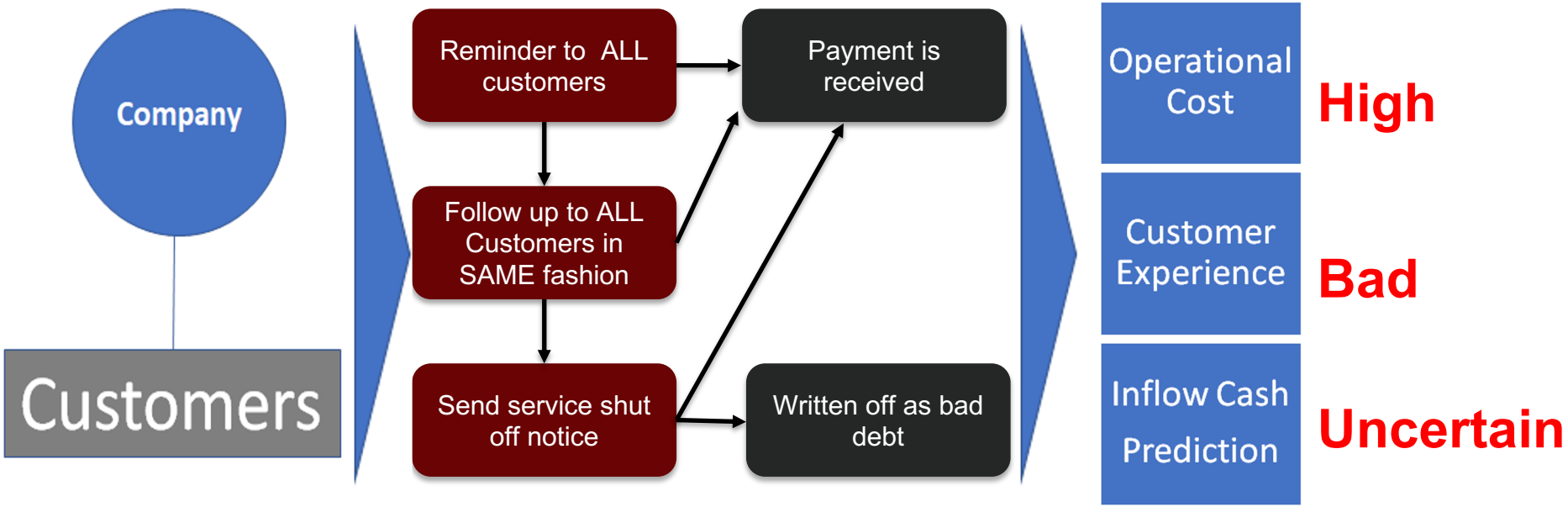
CSCI-P556: Applied Machine Learning Fall 2019

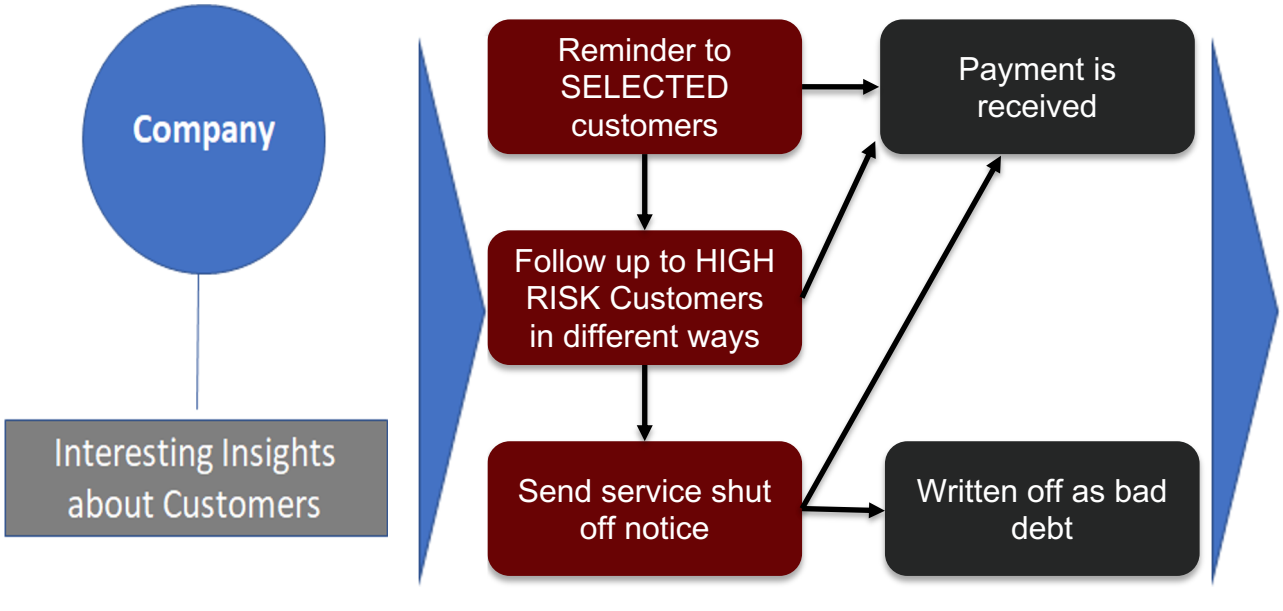
Forecasting Cash Flow and Personalized Customer Experience

Background & Research Questions

The Current Collections Process



The Proposed Collections Process



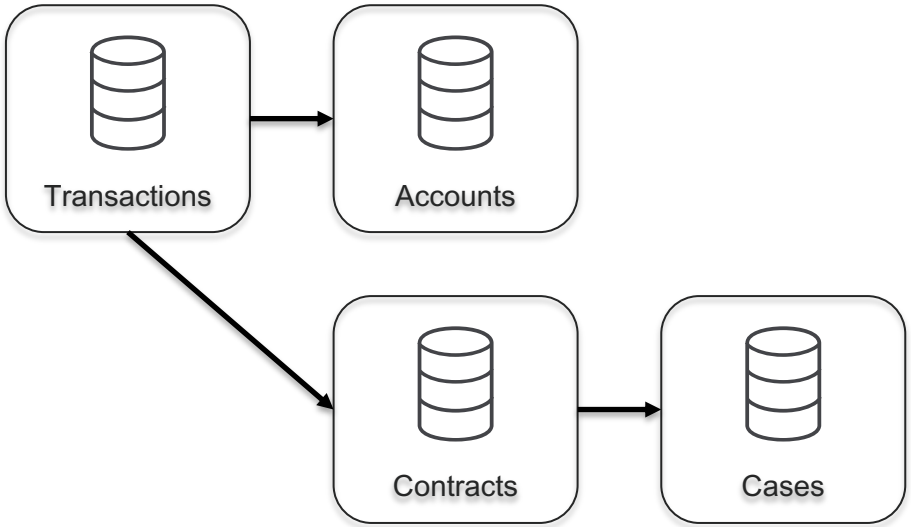
Research Questions

- Can we predict if a customer will pay more than thirty days late?
- Can we predict when a customer will pay?
- Can we classify customers as likely to make a full payment, partial payment or no payment?

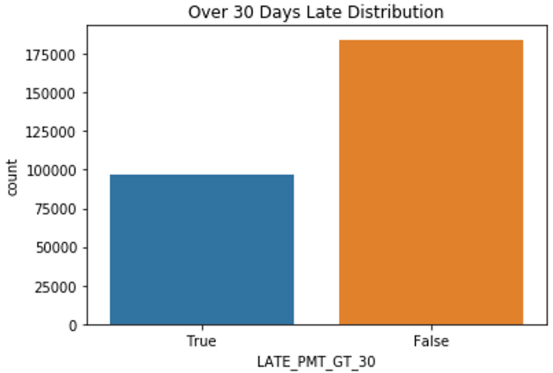
The Data

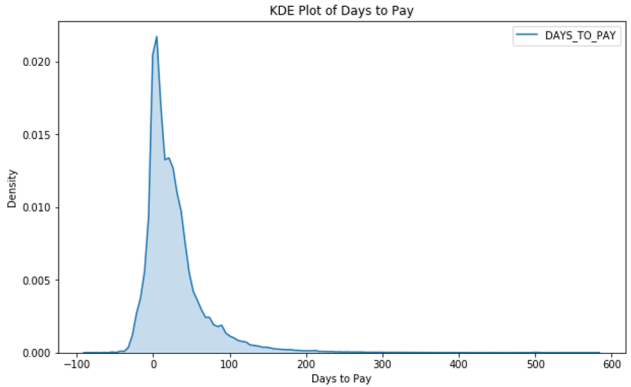
Data Collection

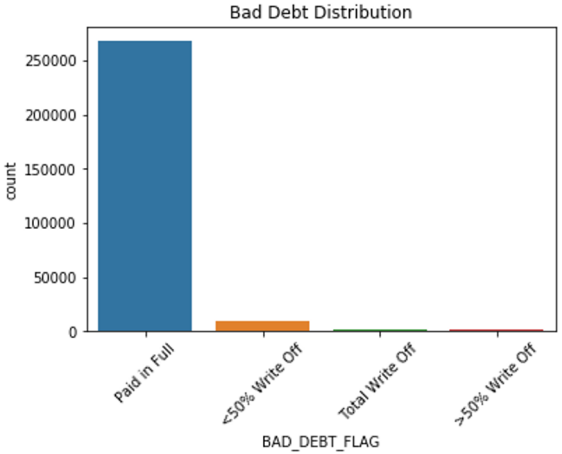
- Base dataset filtered from ~12.6M observations down to ~305k
- Base datasets contain ~689 features. Train set filtered down to 37
- The dataset contains observations from 2018 and 2019



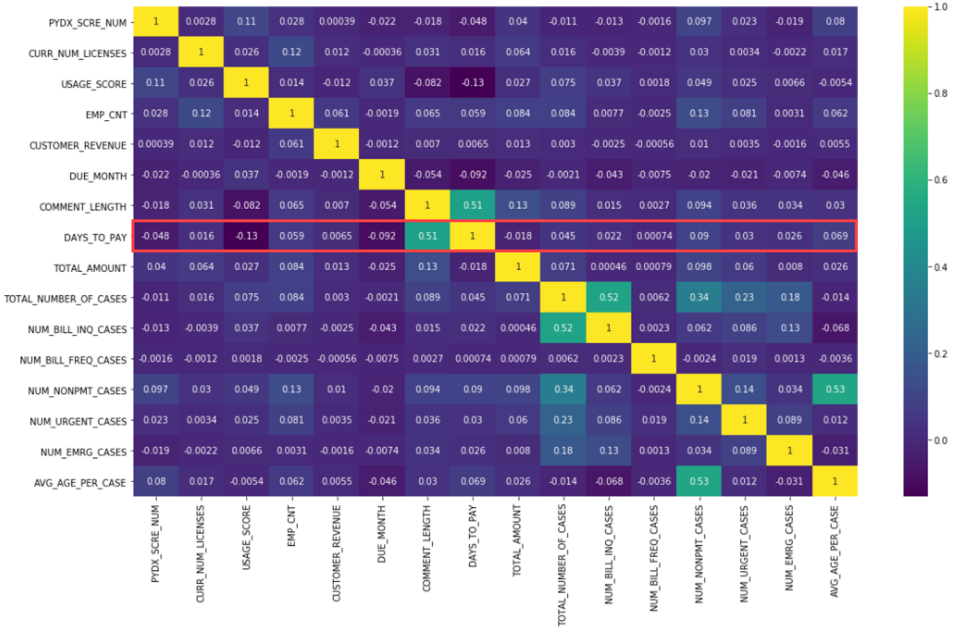
Distribution of “Y” values





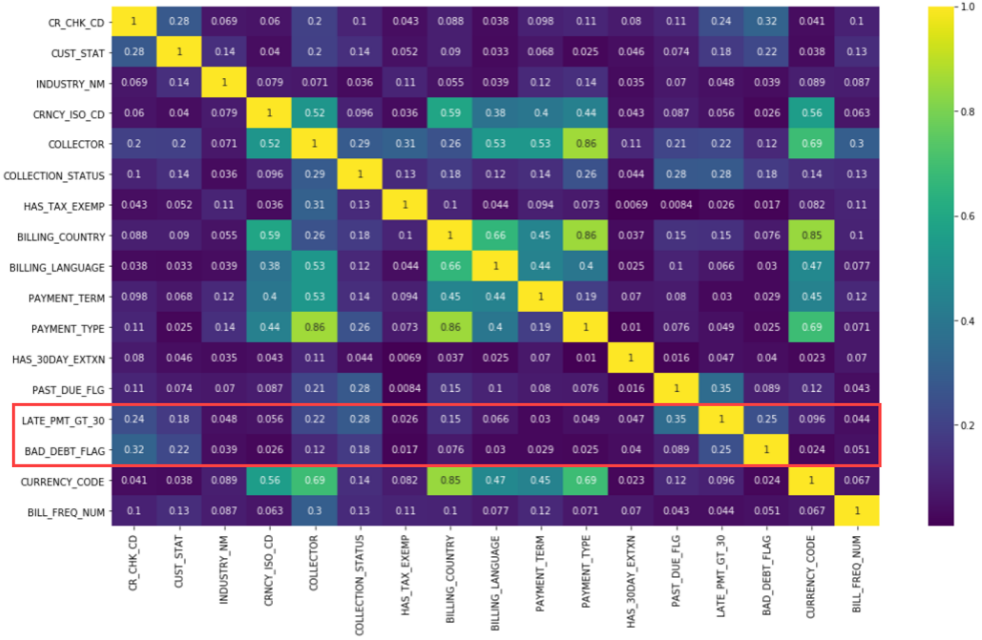


Numeric Variable Correlation



- Usage score has the highest negative correlation (-0.13)
- The variable with the highest correlation was the comment length (0.51)
- No other numeric variables had a correlation higher than 0.1 or less than -0.1

Nominal Variable Correlation

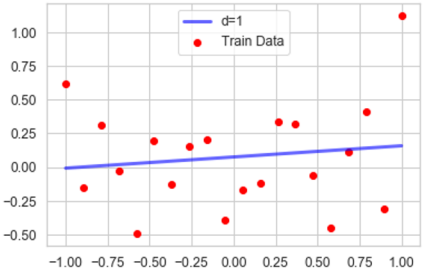


- Associations between nominal variables plotted using Cramer's V statistic
- The variables with the highest association were the collection status (0.28) and credit decision (0.24)
- Correlation matrix of One-hot-encoded collection status shows that a status of "Non-Responsive" has a correlation of 0.22

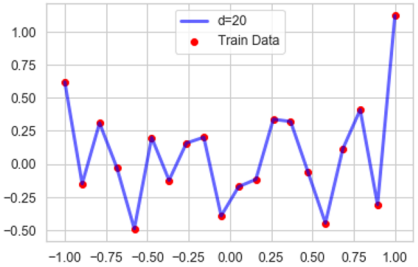
Methods & Results

General Research Method

Simple



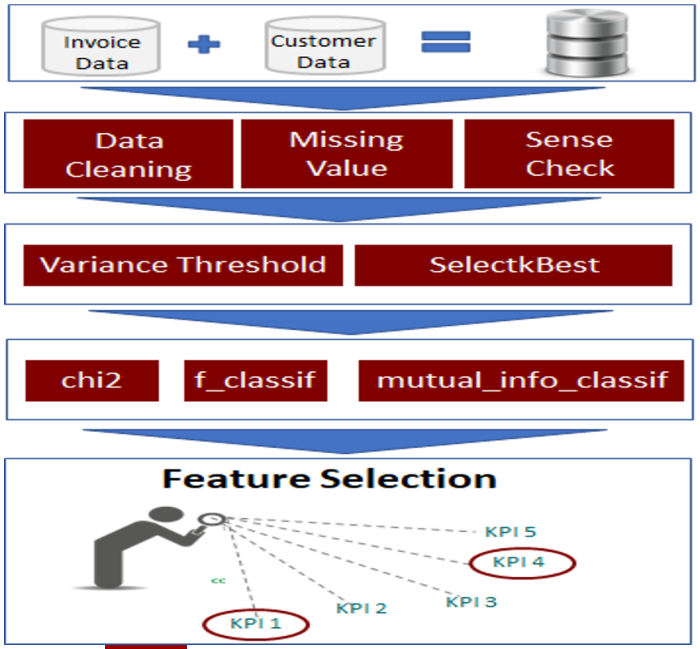
Complex

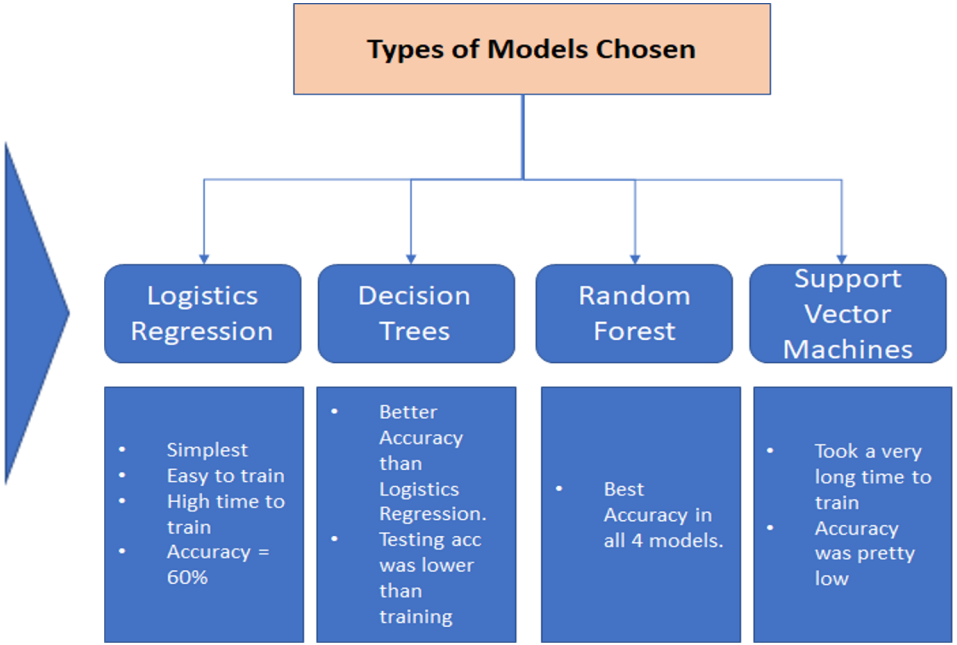


Question 1

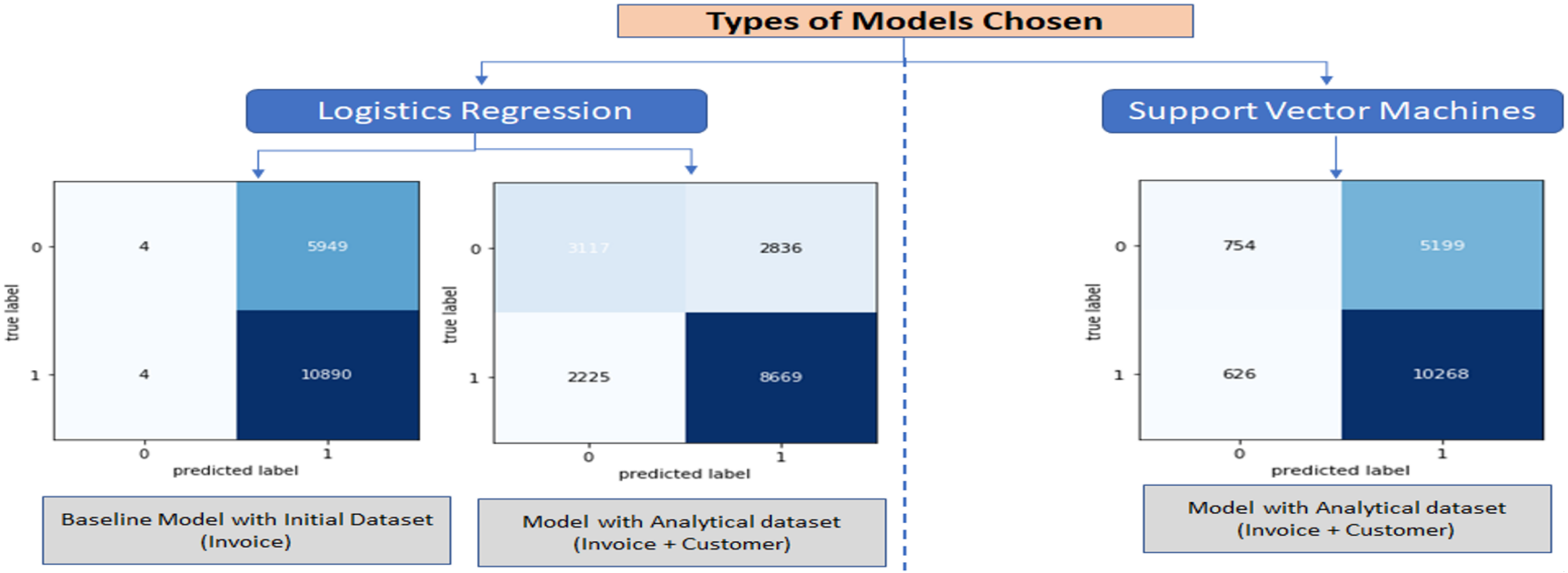
Can we predict if a customer will pay more than 30 days late?

Data Pre-processing and Types of Models

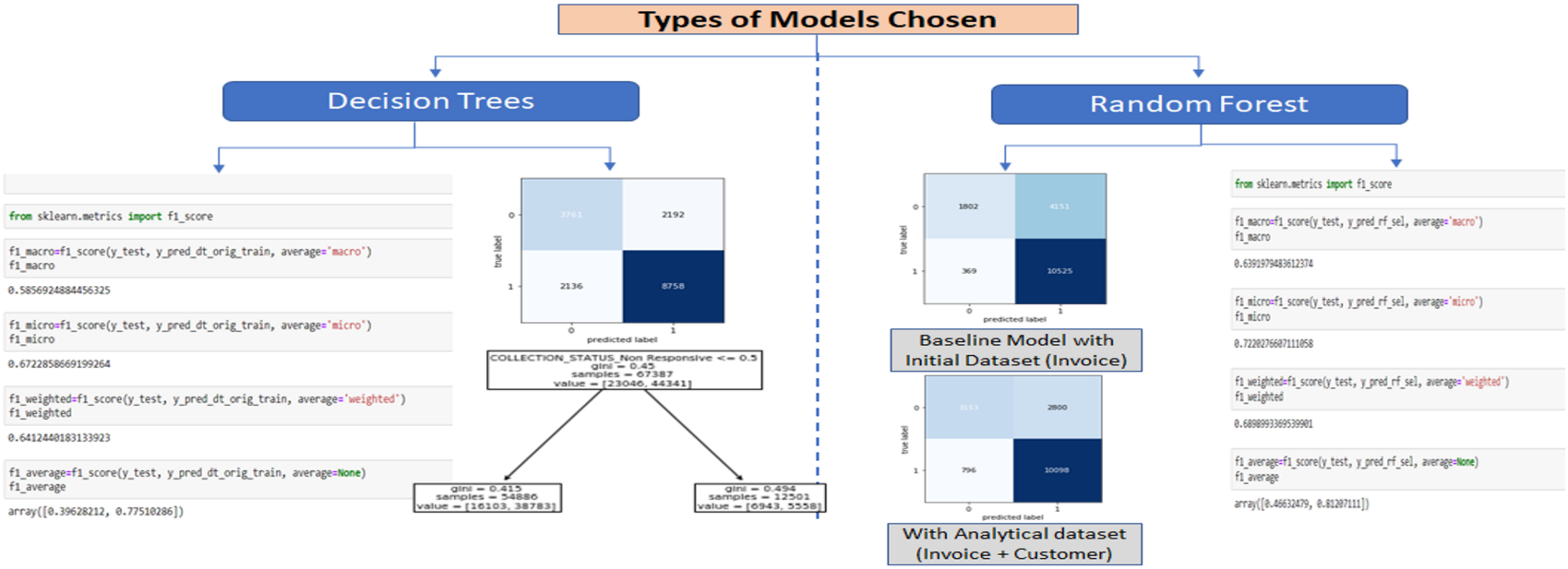




Logistics Regression & SVM



Decision Trees and Random forest



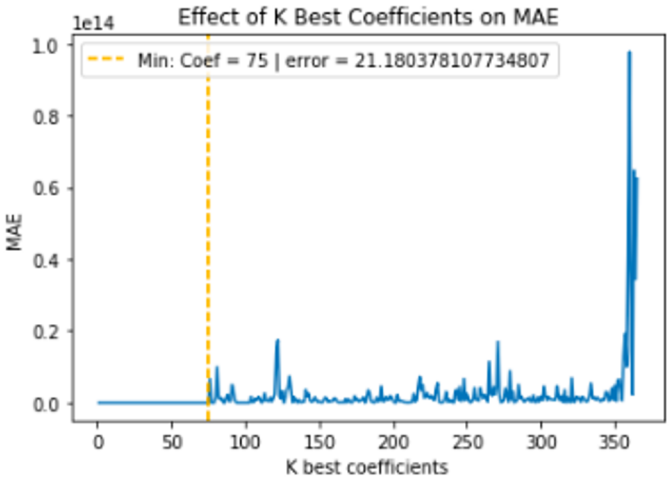
Question 2

Can we predict when a customer will pay?

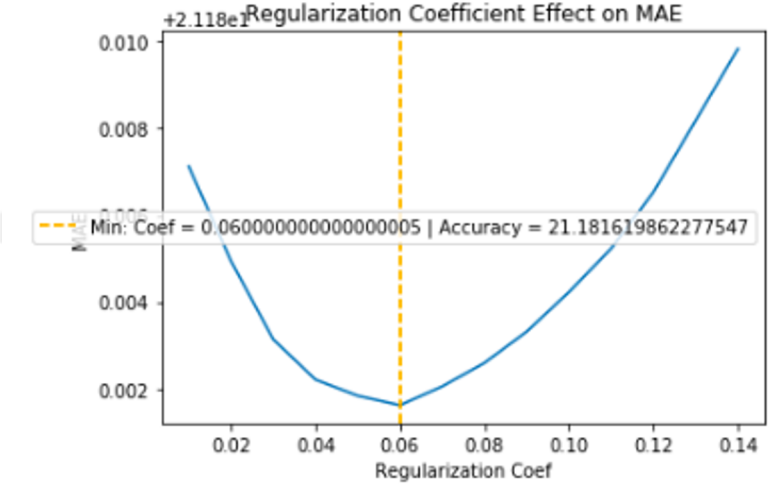
Baseline Model: Linear Regression

Model Optimization

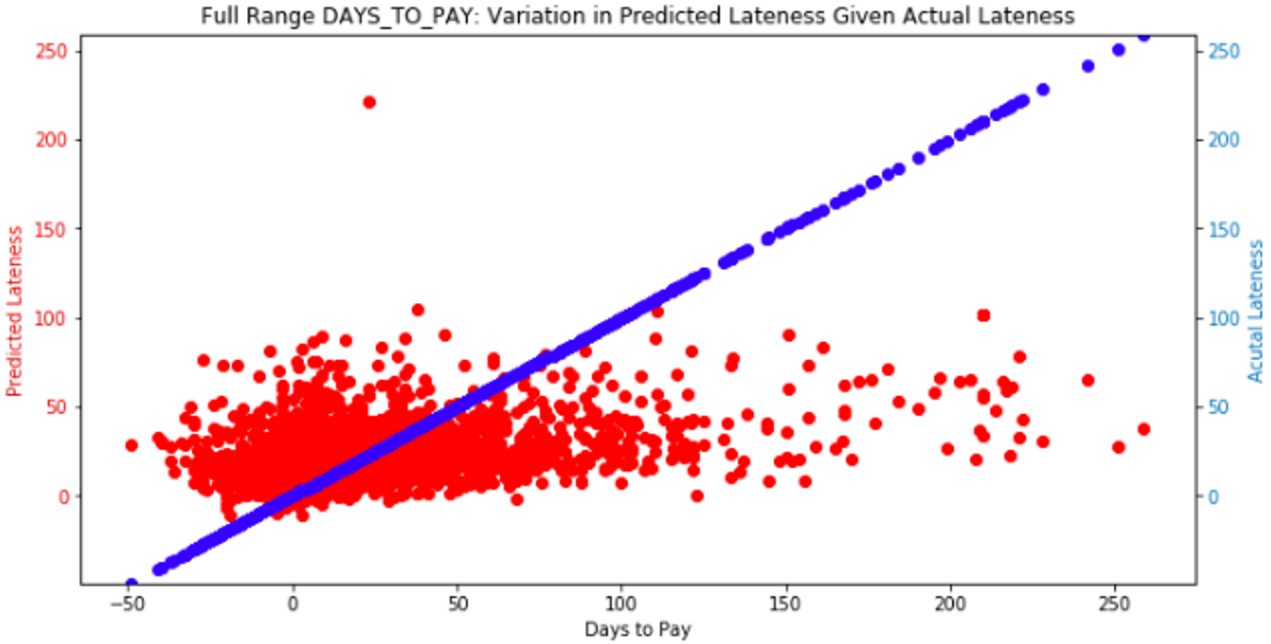
Feature Selection



Ridge Regularization

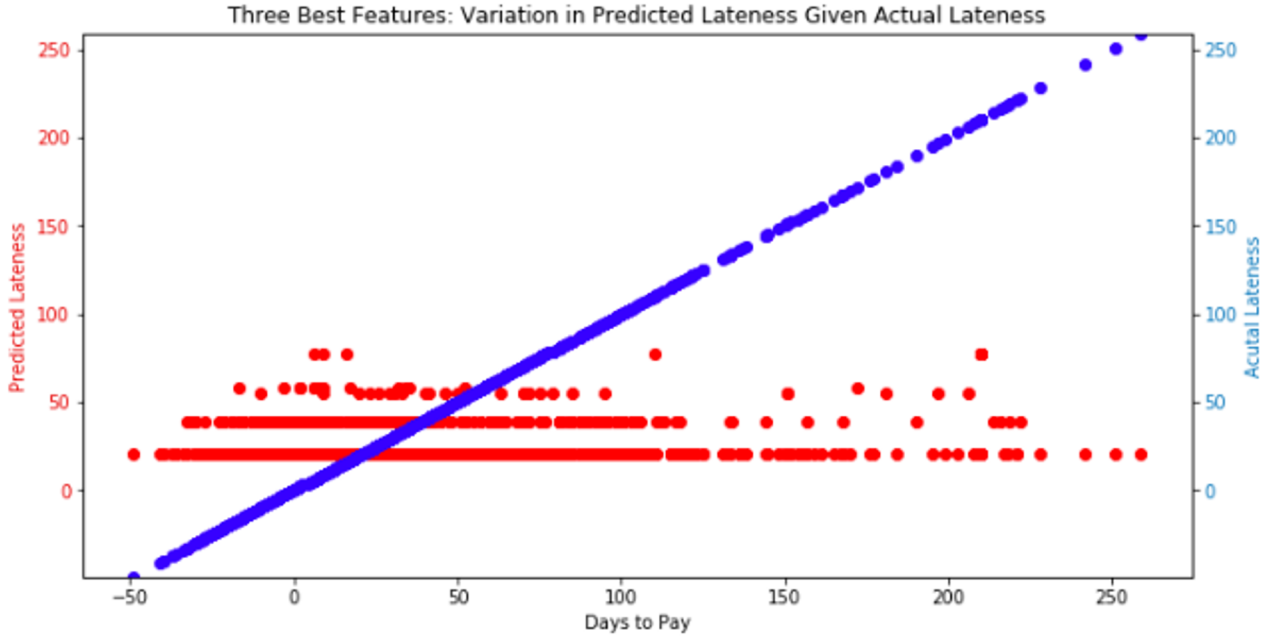


Model Performance



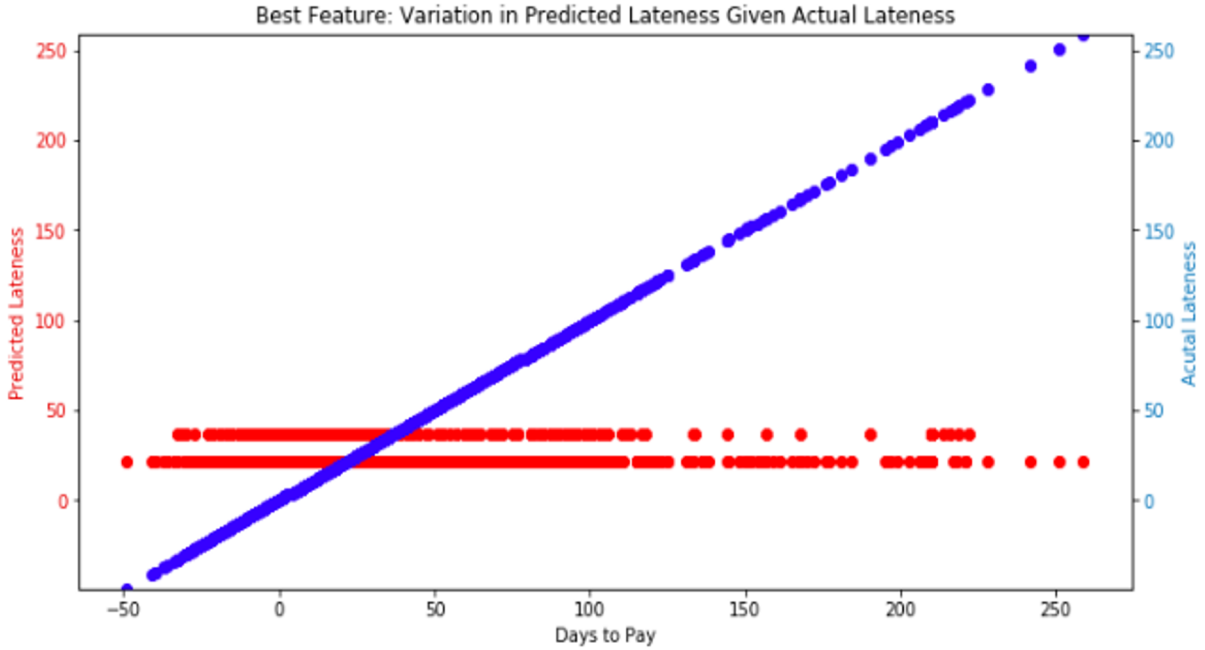
Test error:
22.43 Days

Best Three Features Performance



Test error:
23.96 Days

Best Single Feature Performance



Test error:
23.96 Days

Best Feature Description

Feature Name	Feature Explanation	Feature Effect on Days to Pay
`CR_CHK_CD_Credit Hold`	Boolean Feature based on one-hot encoding of the credit category of the customer. In a credit hold, the customer has hold placed on their credit.	Highest
`COLLECTOR_00530000003iMvZAAU`	Identification number of collector	Second Highest
`COLLECTION_STATUS_Red Account`	Boolean feature based on whether the customer is deemed unhappy or likely to leave.	Third Highest

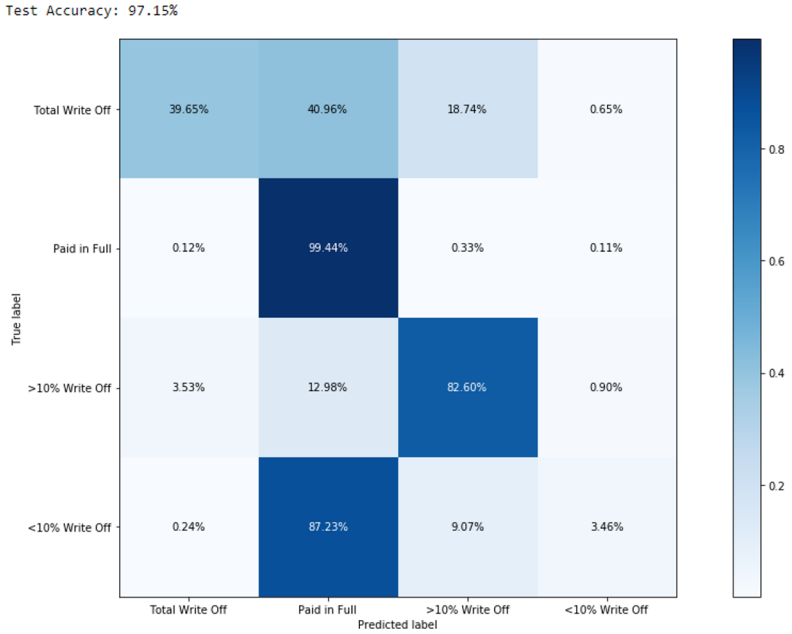
Model Results

Model	Validation Error (Mean Absolute Error)	Test Error (Mean Absolute Error)
Linear Model (all features excluding late payment signifiers)	21.182 days	22.429
Linear Model with only **`CR_CHK_CD_Credit Hold`** (Best Feature from `SelectKBest`)	23.115 days	24.200 days
Linear Model with **`CR_CHK_CD_Credit Hold`**, **`COLLECTOR_00530000003iMvZAAU`**, and **`COLLECTION_STATUS_Red Account`** (3 Best Features from `SelectKBest`)	22.933 days	23.960 days
Support Vector Regression using `GridSearchCV` (C=1, kernel = RBF)	N/A	22.897 days
Support Vector Regression using `GridSearchCV` (C=10, kernel = RBF)	N/A	22.916 days

Question 3

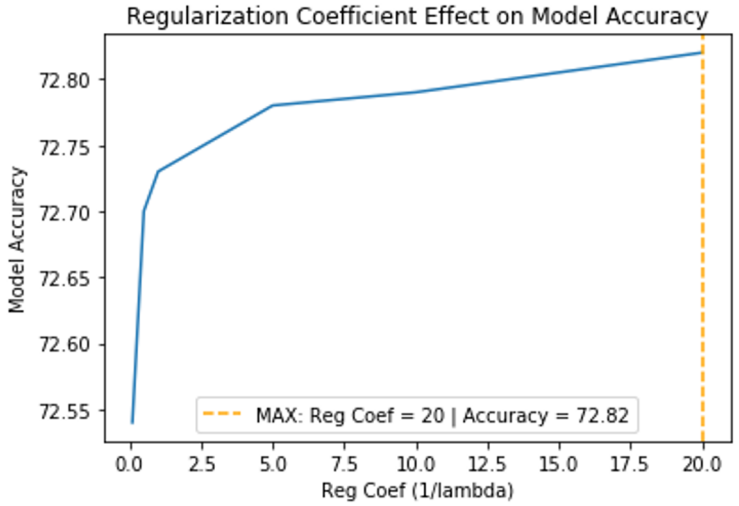
Can we classify customers as likely to make a full payment, partial payment or no payment?

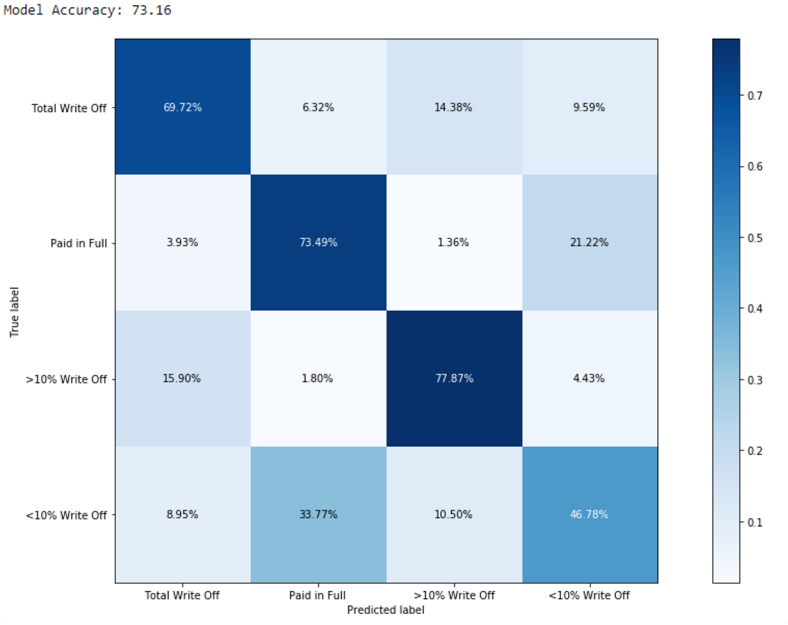
Imbalanced Dataset Results



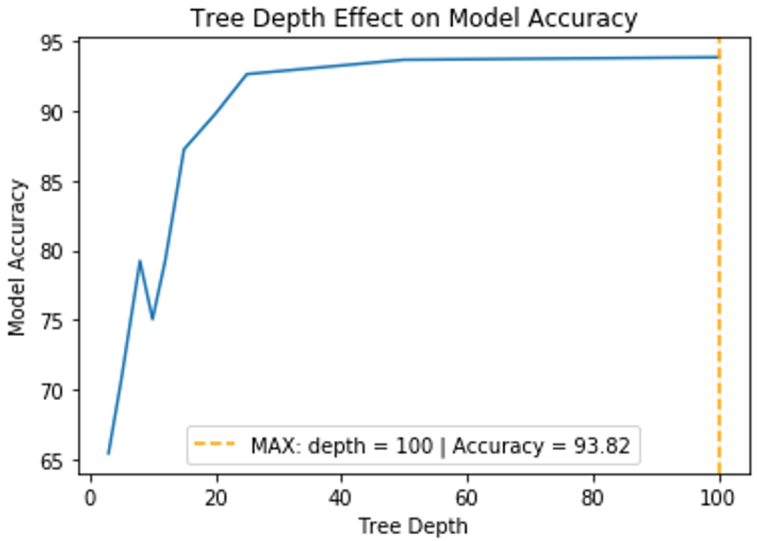
- Overall accuracy = ~97%
- Model was not very good at predicting full write off (39.6%) but predicted paid in full correctly 99% of the time.
- Balanced dataset using SMOTE oversampling technique

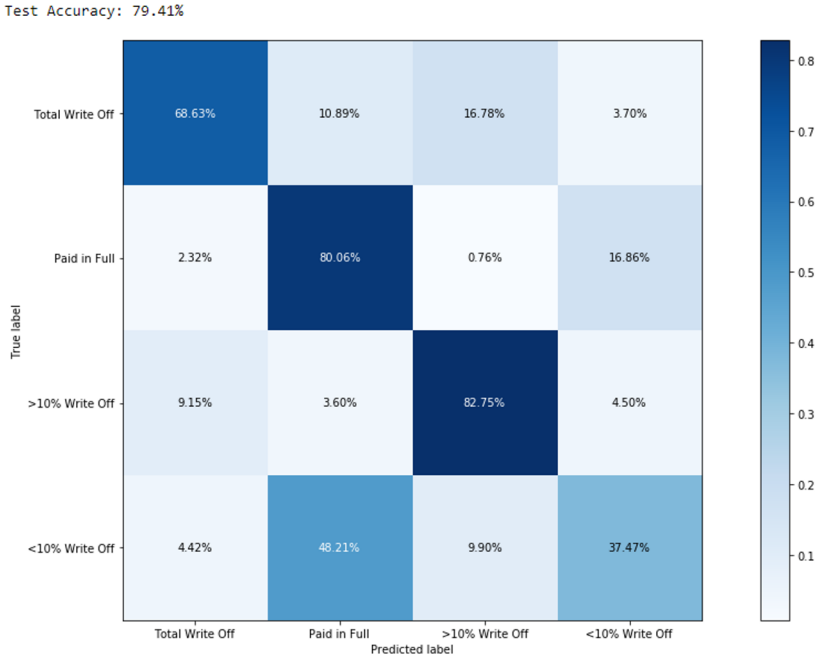
Baseline Model: Multiclass Logistic Regression





Decision Tree





Conclusions

Summary and Next Steps

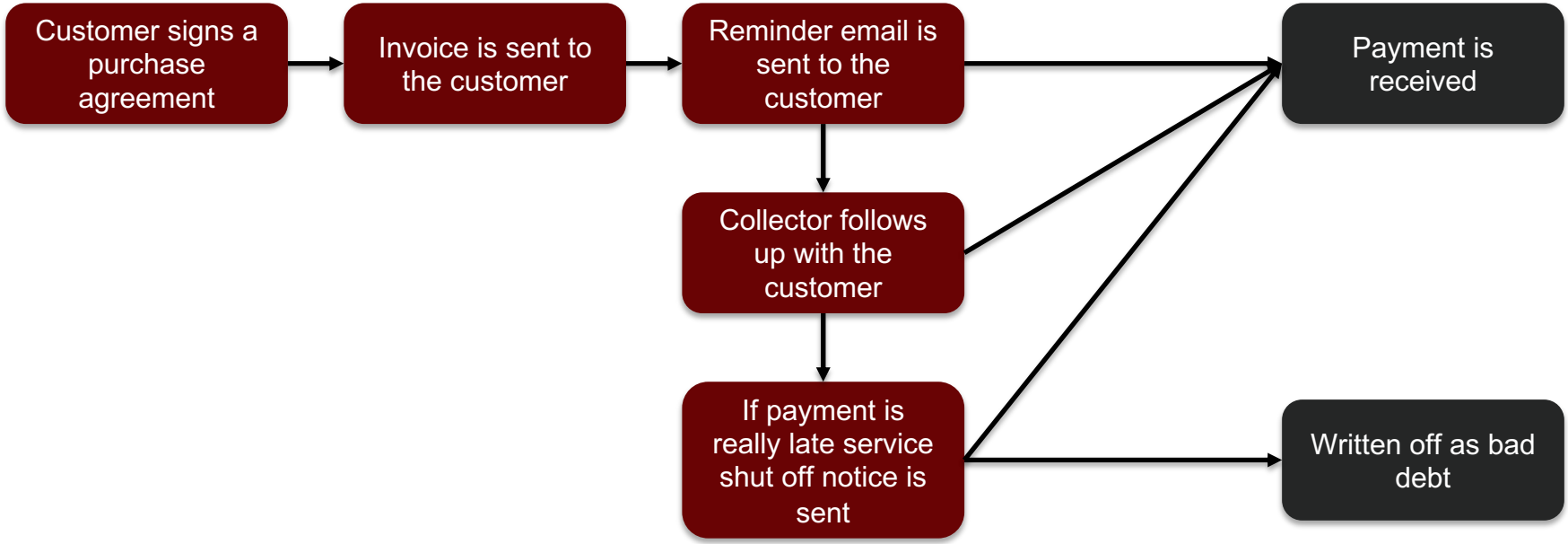
- More Data for better accuracy.
- Experiment with Feature Selection and Engineering.
- Add more features, such as Customer's past records to analyze his payment behaviour.
- Try new models.
- Get data for different industries/ companies and compare results.

Thank You!

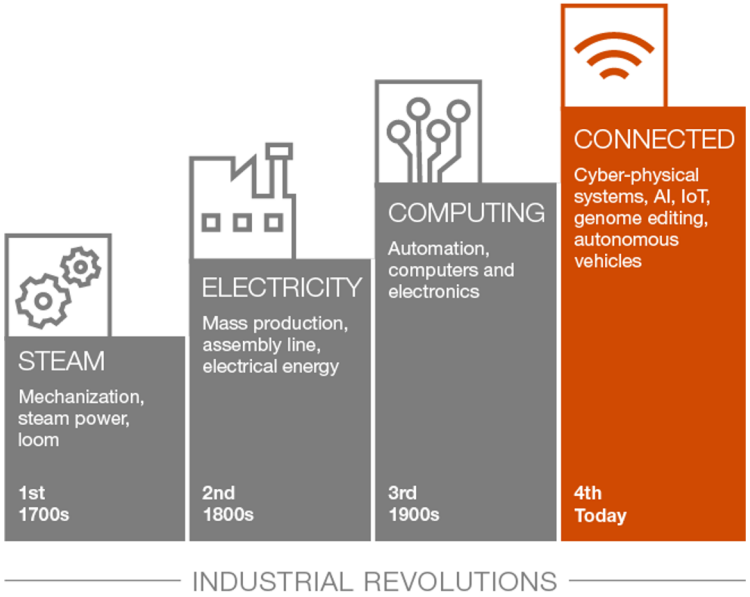
INDIANA UNIVERSITY BLOOMINGTON

FULFILLING *the* PROMISE

The Collections Process



Increasing Stakeholder Expectations



Source: https://www.pwc.com/us/en/library/4ir-ready.html

Research Questions

1. Can we predict if a customer will pay more than thirty days late?
2. Can we predict when a customer will pay?
3. Can we classify customers as likely to make a full payment, partial payment or no payment?

Impact of Improving the Process

4 Different Models and their Results