Executive Summary

Purpose

This document provides an overview of the loan valuation model used to score and rank loans downloaded directly from Lending Club's API. I provide information on general estimation techniques as well as the underlying training dataset. The model was developed to showcase my abilities to manage a data science project from end to end, while participating in the The Data Incubator. It is not intended to be the best possible model that could be built on the Lending Club data, which could involve considerably more research, feature engineering, testing of multiple estimators, etc. Lending club notes are highly complex investments, and each individual should fully understand the associated risks prior to investing.

Estimation Data

The Lending Club data are publicly available and consist of roughly 1.2 million loans underwritten and funded since 2007. Lending Club provides both a static dataset as well as a panel dataset with unique loan-month records. The static dataset is roughly 800MB and contains over 100 loan and borrower features that remain unchanged over the life of the loan. The panel dataset is roughly 4.5GB and includes monthly payments, fees and delinquency status for each loan. The data were last refreshed on February 11, 2017. Both datasets and data dictionaries can be found here:

Static: https://www.lendingclub.com/info/download-data.action

Panel: https://www.lendingclub.com/site/additional-statistics

Note that access to the panel dataset is only provided for Lending Club account holders.

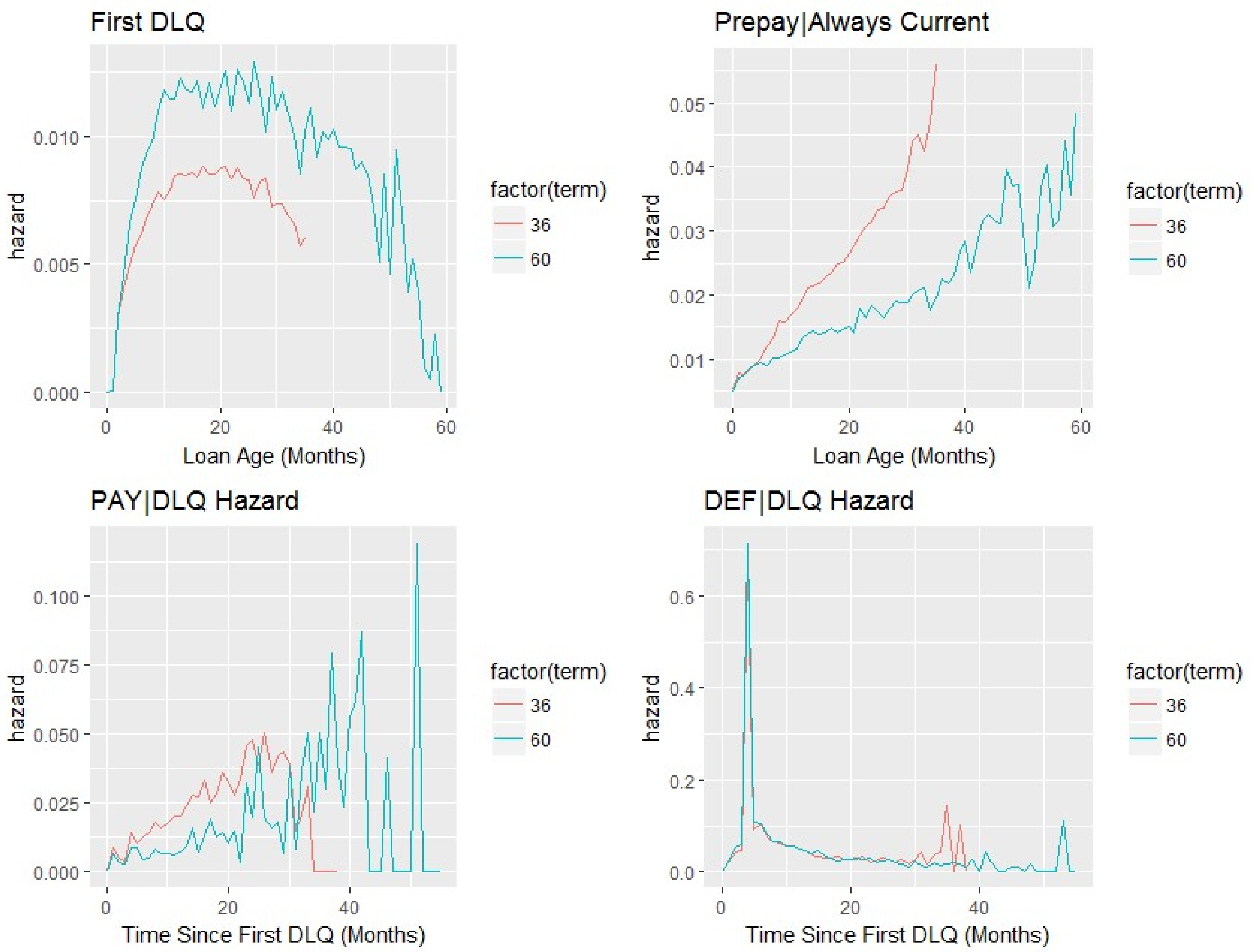
Sampling

Loan termination events such as charge-offs and prepays are, in general, low frequency events. To address this, I oversampled charge-off and prepay events and weighted them appropriately during estimation. This sort of sampling approach also helps to limit CPU, as most of the estimation was performed locally.

Additional filters were applied to the data after some initial data mining. To avoid structural changes in Lending Club's underwriting process, I limited the estimation data to vintages after 2010. Only a couple hundred loans exist in the data with FICO scores below 660, while a large spike in loan counts was noted at the 660 threshold. Loans below 660 were treated as outliers and removed from the data. Similarly, some borrowers had extremely high DTI ratios – some exceeding 1000. These observations were also stripped from the estimation data.

Model Estimation

8 equations were estimated to predict monthly prepayment and charge-off probabilities at the loan level. Separate equations are estimated for 36 month and 60 month loan terms. This is necessary for hazard models (described below) which use time as an explicit feature in predicting a terminating event. After segmenting by term, the strongest predictor for a loan termination event is its current status (i.e. whether the loan is current, delinquent, in default, etc.). While a more complex model might attempt to simulate the transition between all possible stages, I opted for a simplified approach which only uses two conditional states: always current and first time delinquent. Hazards for each state vary significantly across time; therefore, I also segment the model by delinquency status. Actual historical hazards are provided below:



The 4 equations estimated for each term segment are described below.

- **Equation 1, always current to prepay.** The probability of prepay given a loan has never been delinquent.
- **Equation 2, always current to first delinquent.** The probability of first time delinquency given a loan has never been delinquent.
- **Equation 3, first delinquency to prepay.** The probability of prepay, given the loan's first delinquency status.
- **Equation 4, first delinquency to charge-off.** The probability a loan is charged off given the loan's first delinquency status.

Dependent Variable

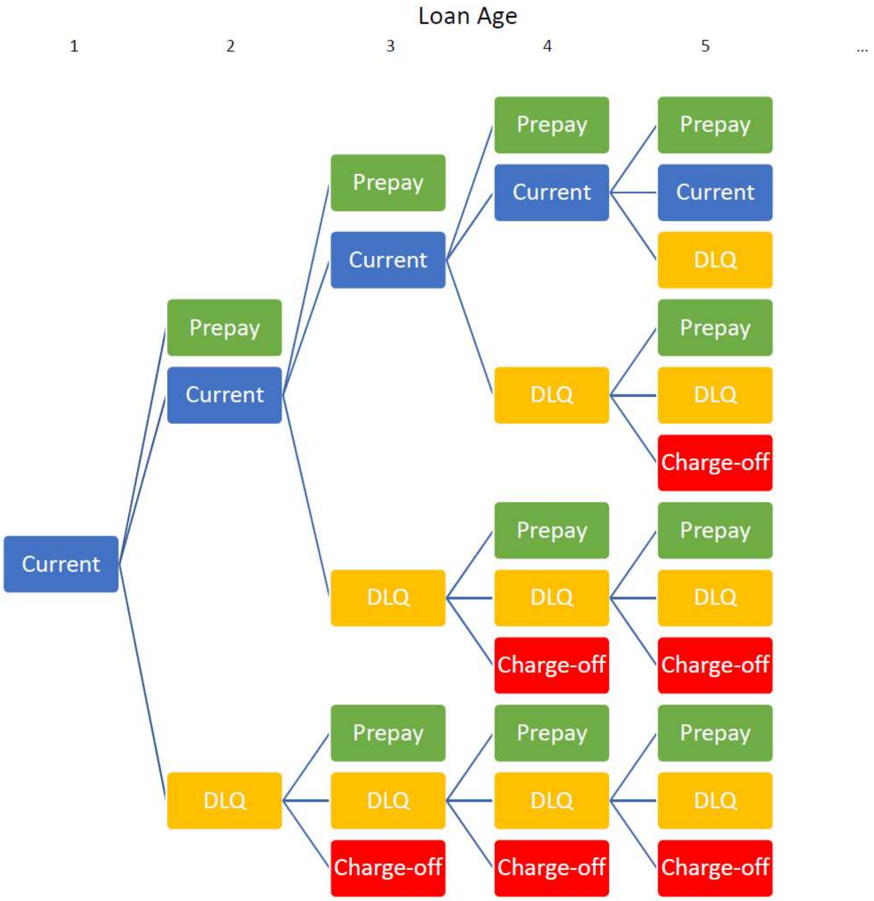
For each of the 4 equations, the dependent variable is the empirical monthly hazard rate. The hazard is defined as the number of events observed over the total number of eligible borrowers. For example, for equation 1, the empirical hazard for month 5 would be calculated as the number of borrowers in month 5 that prepaid divided by the number of borrowers that were not delinquent in months 1 through 4.

Feature Selection

A limited set of significant features were included in the model. Non-linear relationships were fit with linear and cubic splines. With the exception of time, the most important features for first time delinquency are verification status, inquiries, DTI and FICO. Prepayments are generally driven by the borrower's capacity to pay-off the loan early. The most important features for predicting prepayment are DTI, revolving credit utilization and income.

Combining Equations

A visual representation of the model is presented in the following flow chart:



Each loan begins in the current status. In each subsequent month, the loan will either transition to a delinquency status, payoff or remain current. Notice that once a loan moves to a delinquency node, it never returns to a current node. This is because I define “DLQ” as a loan that was ever delinquent. This is a simplifying assumption of the model that limits the branches of the tree. Each transition is given a probability weight, which is later applied to a corresponding cash flow. We apply the following equations to get the monthly predicted hazards for prepay and charge-off.

$$PP_t = \sum_{t=0|i<t}^{T} P(PP_t|Curr) + P(DLQ_i|Curr) * P(PP_t|DLQ_i)$$

$$CO_t = \sum_{t=0|i<t}^{T} P(DLQ_i|Curr) * P(CO_t|DLQ_i)$$

where

- PP = predicted prepayment hazard
- CO = predicted charge-off hazard
- Curr = current status
- DLQ = ever DLQ
- T = loan term (i.e. 36 or 60)
- i/t = monthly time buckets

We can then use these hazards with an amortization schedule to get weighted cash flows in each period. The net present value is simply the sum of all monthly cash flows discounted by the user-defined yield.

Severity

This version of the model does not explicitly measure severity (loss given default). Instead, I make the simplifying assumption that recoveries will be 0 in the event of a charge-off. Preliminary analysis indicated that the data are right censored with an average severity of over 90%. More importantly, apart from the loan amount which was only weakly correlated with recoveries, there were no features that helped classify loans with recoveries. Furthermore, the primary objective of the model is to rank loans, and given there were no loan level features that would differentiate severities at the loan level, a uniform application of 0 severity will have little effect on loan ranking.