Introduction

According to The Federal Reserve in the central bank of the United States[1], the amount of all current residential mortgage loans in United States until first quarter of 2018 is almost USD 15 trillion. Around USD 8.5 trillion of this amount is securitized and sold to the secondary market in the form of Mortgages Backed Securities (MBS). There are three agencies issue and guarantee mortgages in the US:

- the Government National Mortgages Association (Gennie Mae, GNMA),
- the Federal National Mortgages Association (Fannie Mae, FNMA)
- the Federal Home Loan Mortgages Corporation (Freddie Mac, FHLMC).

The combined outstanding debt held by these agencies amounted USD 5.5 trillion. These securities called Mortgage Backed Securities (MBS) which is the most important segment in secondary mortgage market. GNMA securities are default free since they are fully backed by the US government. FNMA and FHLMC are government sponsored agencies but are not default free. GNMA, FNMA and FHLMC securities generally have highly standardized features, trading and settlement mechanisms. The most common mortgage product in the U.S. is the fixed rate fully amortizing loan with a maturity of thirty years. The rate paid on the mortgage is generally fixed for the lifetime of the loan and the loan is fully paid off at maturity. Each period, a constant amount is repaid. The fraction of this amount that is used to repay the principal increases over time since interest payments decrease as the outstanding balance decreases over time.

An important feature of the U.S. Residential Mortgage Market is the agency backed loans that allows penalty-free prepayment of (part of) the principal any time before maturity. Commercial and investment property loans can carry a prepayment penalty, which often consists of a percentage of the remaining principal outstanding. To discourage prepayment behavior, financial institutions in other countries often set a monthly prepayment limit more than which a penalty will be charged.

Standard residential mortgages in the U.S. offer full prepayment flexibility in the sense that entire principal outstanding can be paid off at any time. The option to prepay can be seen as an American option on the mortgage contract. At the discretion of the borrower, the outstanding principal can be returned, relieving the mortgagee from any future contractual obligations on the loan. Although it is possible to form an estimate of the timing of option exercise, such options often are not exercised in an optimal way. The main reason is the diverging ability of borrowers to time optimal exercise of such options as well as other non-rational heterogeneities in borrower behavior.

Aside from the option to prepay the borrower also has the ’option’ to default on the mortgage. The incentive for voluntary default is high when the market value of the property is lower than the value of the mortgage. The default legislation in the US is often targeted as one of the causes of the Financial Crisis.

[1] Source: https://www.federalreserve.gov/data/mortoutstand/current.htm

The most peculiar feature is the fact that in some states[2] mortgage loans are nonrecourse, meaning that if a borrower fails to make contractual mortgage payments, the lender can seize the collateral but has no recourse to any other of the borrower's assets (Gerardi and Hudson, 2010). Voluntary repossession effectively means that a borrower can hand in the keys of the house and be free from any future obligations even though he has enough cash on the bank. Specially when housing prices are low, the existence of negative equity leads to a large amount of so-called "Strategic Defaults". In most countries, the mortgage loan is diminished by the amount paid at the auction. The difference between the market value of the property and the value of the loan, the shortfall, is added to the debt of the borrower. To be alleviated from the entire mortgage debt, borrowers must file for bankruptcy. Voluntary repossessions are not entirely free of costs. They come at the expense of a downgrade in the FICO credit score. FICO scores are used by many US banks to assess the creditworthiness
of individuals. The score is designed to measure the risk of default by considering a range of factors in a person's financial history, such as payment history, size of the debt burden and length of credit history. Lenders are in the position to offer different terms and conditions to customers depending on their credit rating.

The mortgages sold by mortgagees contain implicit options. Prepayment is one of these options. The mortgagor exchanges the unpaid principal balance on the mortgage for a release from further obligations (Deng, 1997). Default is another optionality in a mortgage contact. The property is sold in exchange for elimination of future mortgage obligations. The difficulty with quantifying such options is that mortgagors do not exercise these options efficiently. Moreover, mortgagors' behavior is heterogeneous and cannot be represented by a typical borrower (Deng et al., 2005). This risk that the behavior of the mortgagor deviates from what is expected from a purely financial standpoint is called behavioral risk (Bissiri and Cogo, 2014). The application of this risk to early unscheduled return of the principal on a mortgage is defined as prepayment risk. A formal definition is given in Kolbe (2002) who defines prepayments as: (contractually permitted) notional cash flows which occur earlier or later than expected, deviating from the anticipated call or put policy of the counterparty in a financial contract' (Kolbe, 2008), p.21. Henceforth, Kolbe defines prepayment risk as the risk resulting from these cash flow deviations.

Risk of early mortgage termination makes the duration of a portfolio of mortgages stochastic and in turn has implications on the refinancing policy of the lender (Jacobs et al., 2005). The implicit options embedded in mortgages may or may not be exercised in response to market changes, which in turn leads to significant liquidity risk and interest rate risk for the credit providing institution (Consalvi and di Freca, 2010). Interest rate risk in general arises when there is a mismatch in the fixation of the interest rates paid and received by the bank (Perry et al., 2001). To fund the mortgage, financial institutions attract resources from elsewhere, on which a certain agreed upon rate must be paid over a fixed period. If the mortgagee decides to prepay (part of) the principal at any time before its original maturity, the mortgagee will have to find an alternative use for these funds. If market interest rates have fallen since the origination of the mortgage the bank will incur a loss. On the other hand, uncertainty in the maturity profile of loans subjective to prepayment has a considerable impact on the liquidity representation of the

[2] The following states are classified as nonrecourse states: Alaska, Arizona, California, Iowa, Minnesota, Montana, North Carolina, North Dakota, Oregon, Washington and Wisconsin (Gerardi and Hudson, 2010).

bank ‘s profile (Consalvi and Koning, and E. Sterken, 2010).
An incorrect evaluation of this profile exposes financial institutions to the risk of overestimating future liquidity requirements (overfunding) as well as to the risk of increased long-term liquidity costs (Bissiri and Cogo, 2014). To incorporate the risk of prepayment, financial institutions generally include a charge in their mortgage pricing. This risk is however not priced in completely as a mortgagor can prepay any time while the spread to account for this risk is received monthly across the life of a mortgage (Vasconcelos, 2010). To discourage prepayment behavior of mortgagees, banks usually charge a prepayment penalty. This penalty is mostly equal to the present value of the difference in monthly interest payments between the mortgage and of a newly originated mortgage with the same characteristics (Jacobs et al., 2005).
There exist a multitude of factors that influence the mortgagors’ decision of mortgage termination. These risk drivers can roughly be divided into borrower-specific factors, loan-specific factors and macro-economic factors, (see Table 1).

Category	Explanatory variables
Borrower specific	income, fico score, loan purpose, employment status, Marital status, Debt-to-Income (DTI)
Loan specific	Loan age, loan amount, mortgage rate, insurance, penalty, location, property type, market value of the property, loan-to-value ratio (LTV), Financial Incentive (the spread between existing and prevailing mortgage rates)
Macro-economic	Housing price, market mortgage rate, housing turnover, risk free rate, seasonality

Table 1

Some key features stand out, prepayments often exhibit an S-shaped relation with loan age. This so-called seasoning effect arises since prepayment rates are generally low shortly after origination of the loan and increase as the mortgage matures, the ramp-up period, to finally arrive at a steady state level near the maturity of the mortgage (Charlier and Bussel, 2001), (Jacobs et al., 2005). Secondly, trends in house prices are indicative of the level of activity in the housing and mortgage markets. During house price appreciation, home sales and mortgage originations may increase as the expected return on investment rises (Agency, 2017). Prepayments due to relocation are more prevailing in this case. Conversely, during periods of price depreciation or price uncertainty, home sales and mortgage originations tend to decrease as risk-averse home-buyers who are reluctant to enter the market. In turn, this leads to fewer prepayments due to relocation, and prepayment due to relocation is positively affected by the divorce rate.
Prepayment due to refinancing incentives is by far the most important reason for prepayment. Refinancing can be attractive when the mortgage market rate is below the contractual rate. The contractual mortgage rate paid, and the effective duration of the mortgage are inversely related (Burns, 2010). An important feature of the refinancing incentive is the so-called “burnout effect” (Jacobs et al., 2005). This effect arises due to differences in borrower behavior in a pool of mortgages. If a refinancing incentive occurs (such as drop in mortgage market rates) a wave of prepayments will occur. The borrowers that grasp this opportunity can be deemed the ’fast’, e.g. financially aware, borrowers whereas the remainder of the borrowers in the pool are ’slow’

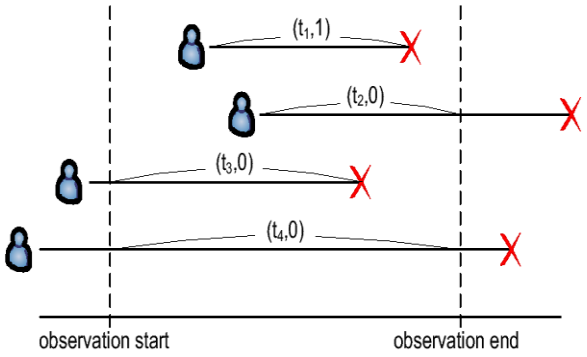
borrowers. Therefore, in the presence of another refinancing opportunity, the pool is expected to be less active or in other words, the pool is burned out (Gonchanov, 2002). Going into strategic default when the market value of the property is lower than the value of the loan is the second most important determinant for prepayment. This is especially true in jurisdictions where mortgage loans are issued on a non-recourse basis. Properly assessing the value of negative equity at each point in time is however a challenging task since the market value of the property is unknown.

Modeling Approaches

Survival Analysis

Survival analysis is a well-established approach to modeling credit loss or prepayment risk for residential mortgage loans. It is also used in both Exploratory Data Analysis (EDA) and predictive modeling. Its application is widespread in biology, social science, and econometrics (Therneau, Grambsch,2000). Survival analysis models the time from mortgage loan origination until the occurrence of pay-down or default, for any given month, when a loan "prepays", and would not be reported on anymore. By definition, leaving the pool is equivalent to "death" or "failure", an analogy in biology or medical researches. Survival analysis uses the statistic for studying the occurrences of the event in certain periods. Then the model parameters are for fitted using the observed events in the historical data. The statistical models consider all the intrinsic and external factors like seasoning and different economic conditions and factors. If we look at prepayment as an event, each loan in our population can exhibit a prepayment behavior in one of the following 4 categories:

a. $(t_1, 1)$ Loan is originated at or after the start of the study and experience the event during the study
b. . $(t_2, 2)$ Loan is originated after starting time and prepayment did not happen until the end of the study, however it happens after observation period.
c. $(t_3, 3)$ Loan originates before starting time and prepayment happens before the end of the study.
d. $(t_4, 4)$ Loan originates before starting time and an event happens after observation time.



Using a big data approach for survival analysis may lead to better performing models.
The **Survival function Surv(y)** is the likelihood that loan will survive and remain current beyond (T). The survivor function plots the natural log of the proportion of a cohort of loans at the beginning (time T_0) that are still alive at t_n . (Schultz, 2016)

Hazard function h(y) is the likelihood that a loan will not survive beyond (T) and the survival function can be estimated in a simple way, (Schultz, 2016)

$$Surv(y) = \frac{Number\ of\ loans\ that\ survived}{number\ of\ loans\ \ at\ risk}, h(y) = \frac{Number\ of\ loans\ that\ prepaid}{number\ of\ loan\ at\ risk}$$

The hazard (mortality) rate at month n is defined as*:*

$Hazard_n = 1 - Surv_n$ The mortality rate in month n is called the "**S**ingle **M**onthly **M**ortality rate" or (SMM). It is the target for our estimation model. The probability that a loan will survive in the next month is our response variable. The Kaplan-Meier (K-M) (1958) is a non-parametric estimator of the *survival function*:

$$\hat{S}(t) = \frac{1}{n}\sum_{i=1}^{n} I\{t_i > t\}$$

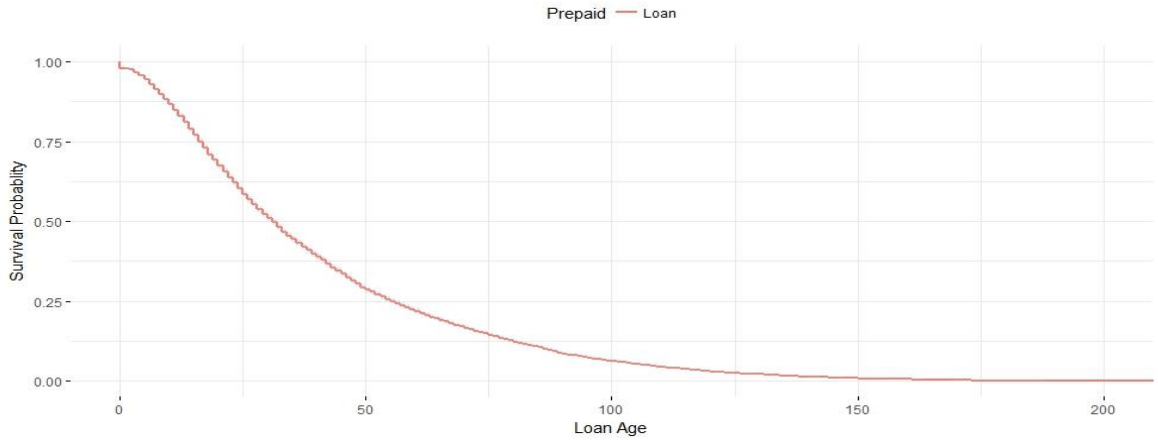
Where i is the indicator function that takes the value 1 if the condition in braces is true, and 0 other wise. The estimator is simply the proportion of loans that are alive at time t. estimator introduced an update of the K-M estimator. It estimates the hazard at each distinct time of death (t_i) as the ratio of the number of deaths to the number of loans that are at risk (Kleinbaum, Klein,2012). The other headers of table pertain to the survival probability for each month, the standard error and the confidence intervals of 95%. So the hazard rate (mortality) (SMM) can be calculated as follows:

$$Hazard_2 = 1 - Survival\ Prob_2 = 1 - 0.9901 = .0099$$

The cumulative survival rate given in the table for the loans at age 6 (months) is 0.9642. It means 96.42 % of the loans under study have survived to age 6. The cumulative survival is equivalent to the MBS pool factor. The MBS pool factor is simply the ratio of the current balance outstanding relative to the original balance of the pool. MBS pool factor reports the percentage loan balance that has survived, after giving consideration to scheduled principal through loan or pool age (n): (The bond market Association,1999)

POOL FACTOR = SURVIVAL FACTOR * AMORTIZED LOAN BALANCE

The plot for total probability of survival for loans throughout their life from January of 1999 to March 2018 (223 month):



If we annualize the SMM we reach an important rate for pool of Residential Mortgage loans so-called conditional prepayment rate.

$$\text{CPR (Conditional Prepayment Rate)} = 1 - (1 - SMM_n)^{12}$$

Multinomial Logistic Model (MNL)

Discrete choice models are models in which the dependent variable is a categorical response variable$Y_{it} = j$, for j = 1,2,3. Using a set of explanatory variables, these models estimate the probability that either one of the categories in the ordinal dependent variable occurs. Two popular form of this model are the Probit and the Logit. The difference between the two lies in assumptions for the distribution of error terms. The probit model assumes a normal distribution, whereas the logit model assumes a logistic distribution. In the context of mortgage continuation and termination, the multinomial logit model (MNL) is the most widely adopted When the dependent variable is a probability, a linear relationship between the dependent and the explanatory variable can't be appropriate. Instead, the MNL model tries to predict the prepayments, assuming that outcome is a discrete variable that shows the status of the loan when the event of prepayment occurs. The multinomial logistic model (MNL) in more details will be defined as:

Let Y_i be random variable that one of the finite number of values, 1,2, …, J.

$$\text{Let } p_{ij} = P(Y_i = j) \; so \; \sum_{j=1}^{J} p_{ij} = 1.$$

For grouped data, Y_{ij} is the number of observations failing into category j,

$$\text{for each } i \text{ Let } n_i = \sum_j Y_{ij}$$

For ungrouped $n_i = 1$ and only and only one of $Y_{i1}, \ldots, Y_{iJ}$ is equal to one and rest are zero.
The Y_{ij}, conditional on the total n_i, follow a *multinomial* distribution:

$$P(Y_{i1} = y_{i1}, \ldots, Y_{iJ} = y_{iJ}) = \frac{n_i}{y_{i1}!, \ldots, y_{iJ}!} p_{i1}^{y_{i1}} \ldots p_{iJ}^{y_{iJ}}$$

About restricting the probability between zero and one like binary response model, there is a need for link function in multinomial model that links the probabilities p_{ij} to the predicators x_i,

with similar idea:

$$\eta_{ij} = x_i^T \beta_j = log \frac{p_{ij}}{p_{i1}} \text{ j= 2 , … , J}$$

It is essential to have $\sum_{j=1}^{J} p_{ij} = 1$,so it is convenient to declare one of the categories as the *baseline*(benchmark), say j=1

So, we get $p_{i1} = 1 - \sum_{j=2}^{J} p_{ij}$ then we have:

$$p_{ij} = \frac{\exp(\eta_{ij})}{1 + \sum_{j=2}^{J} \exp(\eta_{ij})}$$

Then $$P(Y_{i1} = 0) = \frac{1}{1+\sum_{j=2}^{J} \exp(\eta_{ij})}$$

The maximum likelihood of the model is:

$$Ln\ L(\beta_j) = \sum_t \sum_i \sum_j d_{ijt} ln\, P(Y_{it} = j)$$

Where, d_{ijt} is a dummy variable for category of Y_{it}

With noticing that $\eta_{i1} = 0$. It does not matter which category declared as the baseline although some choices may be more convenient for interpretation. We use maximum likelihood and then use standard methods of inference as General Linear Methods.

Data Description

The dataset used in this research consist of Single Family fixed rate mortgage loans reported on monthly bases by the Freddie Mac until mortgage termination or maturity. The total number of

sample's observation is 811,135 observations. Loans have origination from January 1999 to December 2017.The observation window is January 1999 until March 2018.

Dependent Variable (Termination)

The dependent variable used in model is a categorical variable indicating three states of the principle payment scheme. The Three states are:

- Contractual prepayment
- Complete prepayment of the mortgage
- Default

Contractual prepayment is the Benchmark category. It will be shown by $Y_{it} = 0$. It is noticing the loans that paying the loan according the originated contract, their payment is not changed during the study.
Complete Prepayment loans will be shown by $Y_{it} = 1$. This category identifies the loans that voluntarily prepays full remaining outstanding principal. This classification is derived from the dataset by means of a variable indicated by zero balance code. The category $Y_{it} = 2$ denotes Default. Loans comprises foreclosures by an alternative group, for example through a short sale, third party or charge of note sale. In this case borrower is unable to make payment for principle or interest and property can be seized. A repurchases prior to a property disposition and a Real-Estate-Owned (REO) disposition are also classified as default. A REO deposition occurs if the lender becomes the owner of the property after an unsuccessful Foreclosure auction. Table below provided the number of observations for the dependent variable per category for random sample:

Y_{it}	**Description**	**Percentages %**
Cntractual payment (Active)	811135	98.4664
Voluntary prepayment	12203	1.41
Default: 30 days	141	.1711
Default: 60 days	45	.0054
Default: 90 days	244	.0296

Independent Variables

Following table provides a description of explanatory variables for mortgage *i* and time *t* that are chosen from our Freddie Mac Data Set. There are some other complementary variables are added. Macro-economic variables are monthly state level Unemployment rate, state level monthly House Price Index (HPI), different between two-year Treasury and ten-year and Monthly Market mortgage rate. The added rates are Not seasonality adjusted.

Variables	***Descriptions***
$LoanAge_{it}$	The number month since origination of Mortgage

$Curr\_int\_rt_{it}$	shows monthly rate of mortgage note, constant during contract due to fixed-rate of loan
$FICO_i$	Credit Score, this score is reported only in origination
$First\ home\ flog_i$	Indicate whether the property is the first home bought by the borrower
$insurance_i$	The percentage of loss coverage on the loan (between 0 and 55 percent)
DTI_i	Debt to Income (DTI) ratio defined as the sum of the borrower's monthly debt payment divide by the total monthly income used to underwrite the borrower at the origination date
$LoanSize_i$	The Original Loan Balance, Scaled by the mean of sample for modeling purpose
LTV_i	The original Loan_To_Value, ratio defined by the original principle divided by the purchase price of the mortgaged property.
$Penalty_i$	Dummy variable indicates whether the borrower obliged to pay a penalty for unscheduled return of principle
$OccupanyOwner_i$	Variable indicating that house is owner occupied, investment property or second home
$PropertyType_i$	Denote whether the property is a condominium (CO), Planned Unit Development (PUD), Cooperative Share (CS), Manufactured Home (MH) or Single-Family Home (SFH)
$Purpose_i$	Indicates whether the mortgages are a cash-out, refinance or purchase (exclusively)
$State_i$	Indicating the Geographical location, the state that property located
$Uenmploy\ rate_{it}$	Unemployment Rate adjusted base on location(state) and Date
HPI_i	House price Index, adjusted base on location(state) and Date
$Treasury\ rate\ Diff_{it}$	Difference between Swap rate on 2yrs and 10 yrs. Treasury bond base on location(state) and Date (seasonal not adjusted)
$Incent_{it}$	Difference between Mortgage note rate and market mortgages rate
Channel (Disclosure type)	Indicating whether any kind of third party involved in origination. Correspondence (C), Broker (B), Retail(R), TPO not Specified (T)
Number of Borrowers	Dummy variable indicates whether number of borrowers is Single or multiple

Variables	Mean	Standard Deviation	Max	Min
Incentive	0.6539	1.082283	6.5600	-7.160
Cnt_borroDum	0.5767	0.4940826	1	0
Flag_fthbDum	0.1063	0.3082179	1	0
Delq-stsDum	0.04011	0.1962295	1	0
Loan_PurposeN	0.3051	0.4604432	1	0
Loan_PurposeC	0.2827	0.4503345	1	0
Loan_age < 12 month	5.686	3.46483	11	0
12<Loan_age <36 month	23	6.588639	35	13
36<Loan_age <72 month	51.26	9.841622	71	36
72<Loan_age <195month	101.9	24.90312	194	72
Loan_age <195 month	205.9	6.995468	223	195
Current_LTV	46.31	34.89809	169.50	0.00
SATO	-8.52000	0.7439354	- 3.22000	
Fico	733.3	55.53286	831.00	300.0
Channal C	0.1344	0.3410507	1	0
Channel B	0.04687	0.2113562	1	0
Channel T	0.3407	0.4739422	1	0
State	0.040100	0.03251349	0.123300	0.001721
Seasonality	1.0030	0.1176565	1.1540	0.7848
Unemployment	6.472	2.207262	15.400	1.800
LTV	71.76	16.93268	100	7.00
T10Y2YM	1.576	0.7976936	2.83	-0.410

Dealing with Categorical Variable

However, cell coding is the most frequent method for dealing with Categorical variables. There are some other ways can help us to have better view of influential factors. One of Obstacles of using binary coding for categorical values, is categories with many members. In our study state of mortgages is categorical variable and it has fifty category types. In result, we need at least 49 columns in binary coding method for none-singular. Then we will face big sparse matrix that reduce model performance and cause memory size problem in big data.

The other obstacle is the influence of binary coding in target variable. The example of unsophistication is when a variable with one member that has much lower frequency, but important correlation with response value. Therefore, it will get low significance. In the type of modeling that target values are not continuous this value with binary coding method can be underestimated and have weak performance, however it has more prominent role.

By considering above mentioned arguments, A three step method has been used for choosing the best method or importing the categorical values. At first step, all possible methods applied on each variable. The three prominent methods are: the frequency of each members of category, using the probability of occurrence and probability that response value be influenced by that value by using cross table.

At step two, cross validate the each of the methods for each categorical value, for example if there are 5 categorical value and four methods, twenty cross validations executed. In each time AIC Score (H. Akaike,1971)[3] and residual deviations were recorded then lowest values for each category will be selected as a representation for that variable.

Then, we need to check for mutual independence of them. The lowest independence or no dependence must be chosen. The test is used is Pearson's $\chi^2 test$.

At Last step, we use the survived variables in Cox Proportional model. Since the Cox Proportional model is simpler than the multinomial model. In this modeling, the assumption is having Poisson distribution of General linear models (A.Dobson, 2008). In action, modeling for each age-group of loans separately, which is called "Survival Object". Therefore, loan with same age will be modeled in same time and we can expect that normal environment that we can use Wald test and p-value for examining the significance of each variable. Not significant variables will be removed with considering the model Wald Test and p-value and AIC, if the results where hard to decide we need to check other forms of presentation of variable in model that helps about the leaving variable.

Below the variables that is been considered in modeling is been shown.

---

[3] Let $\hat{L}$ be the maximum value of the likelihood for the model. Then the AIC value of the model is

$$AIC = 2k - 2\ln(\hat{L})$$

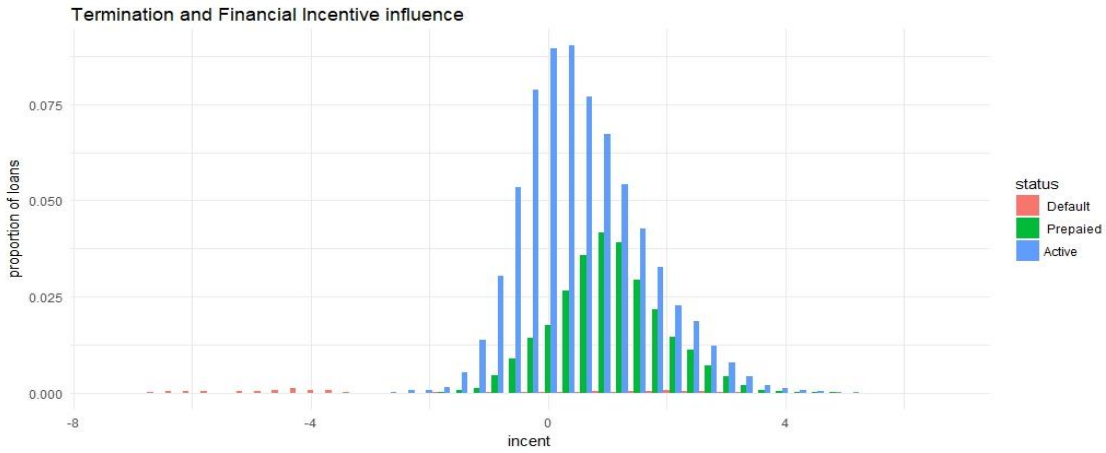
Given the set of candidate models for the data, the preferred model is the one with minimum AIC Value

Refinancing incentive

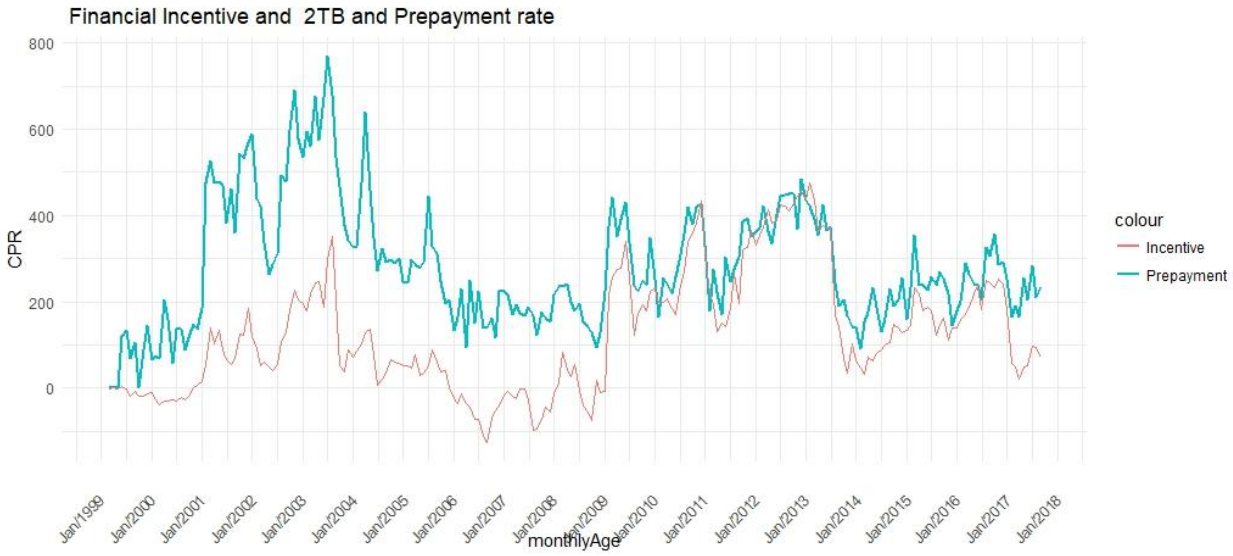
The borrower's incentive to refinance refers to the option to prepay in positive economic condition due to declining current market mortgage rates. However, there are many ways to measure this important variable. In this study we define measure refinancing incentive as the difference between each mortgage note rate and the prevailing Market mortgages rate:

Incentive spread = Note rate – prevailing market Mortgages rate

Market rate is referred to 30-year average market rate that is been reported by Freddie Mac web site. As can be seen in the histogram, the number of prepaid loans (green bars) in positive side of incentive is much higher that the negative side of it. It appears that this variable has a normal distribution. It is balanced and not appear to be skewed.

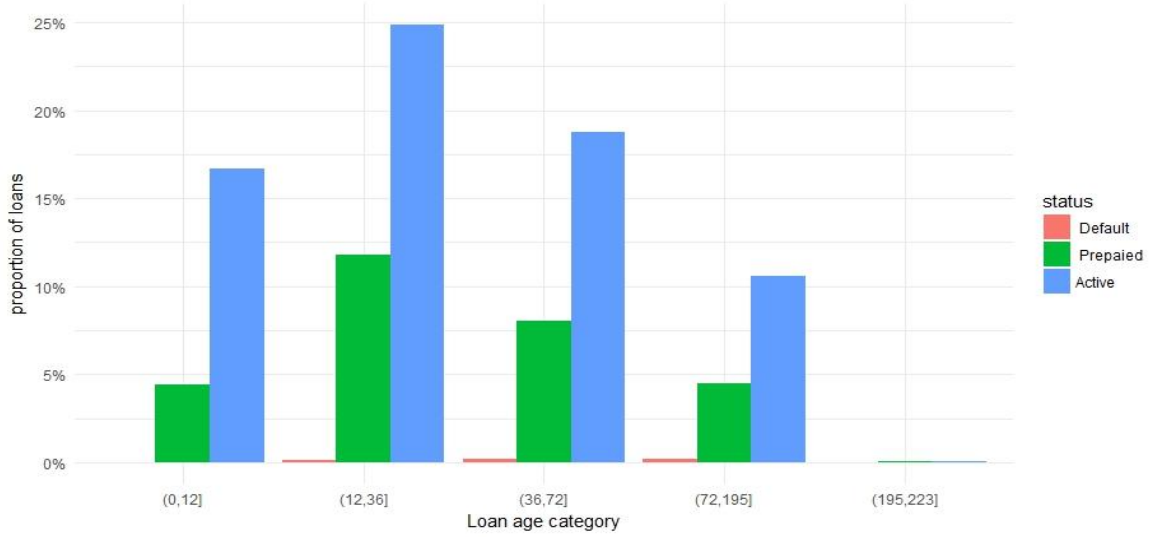


The graph below shows how significantly the financial incentive influences in the prepayment of loans in our portfolio. Comparing the trend in both and picks in both, it is easy to see when there is a stronger Financial incentive, there is a raise in the CPR. Since There are other prominent risk drivers that influence borrower's decision for prepayment, we need to add more variable to have better modeling perspective.



Loan age

Loan age is one of the most crucial factors in analyzing the prepayment of residential mortgage loans. During the life of residential mortgages loans, prepayment behavior will not stay the same, as the loan ages. During the first year, the borrower may like to refinance for other reasons than the economic incentive. The first one is credit cure, after paying for around 7 consecutive periods without delinquency borrowers can refinance and have better rate, meanwhile he does not need much documentation, so the process is easier to precede.

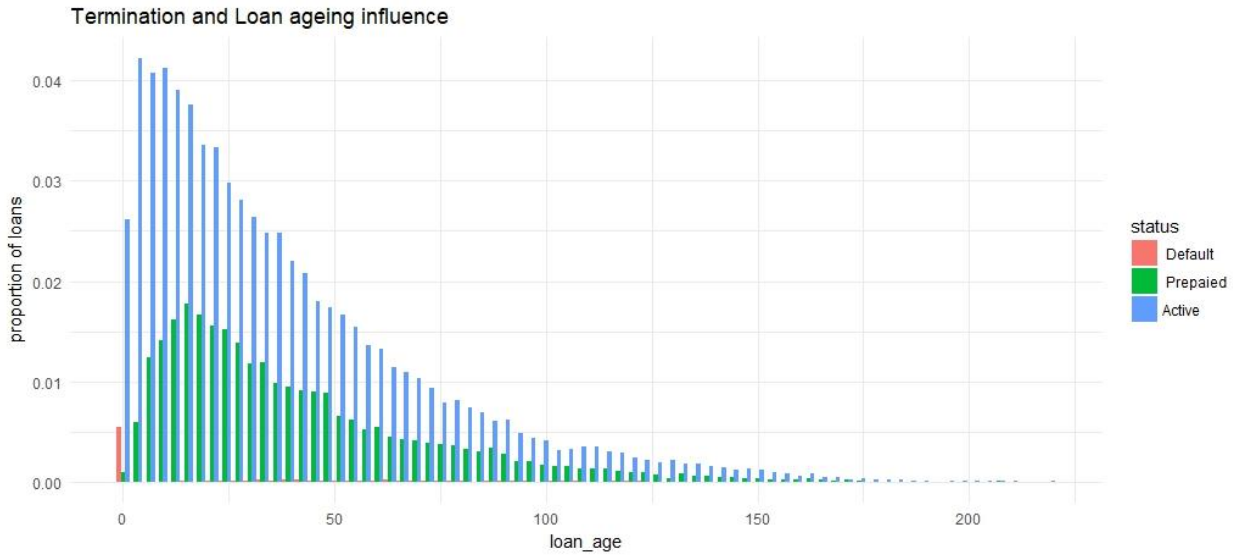


For the loans between 2 and 3 years old all the above-mentioned reasons are possible albeit with lower speed and less sensitivity to incentive spread. Other factors play a stronger influence. The factors include upward trending house prices leading to cashing out the loan by taking another loan with better coupon rate, because of lower loan to house value (LTV) ratio, better credit score etc.

The loans that have lost the opportunity to prepay in early yearly become **fairly seasoned** loans. Loans in this age group may have experienced delinquency or missed the opportunities for many other reasons. However, the refinancing will not be easy for such borrowers in the future.

There are other factors that can drive prepayment other than the ones we just listed:

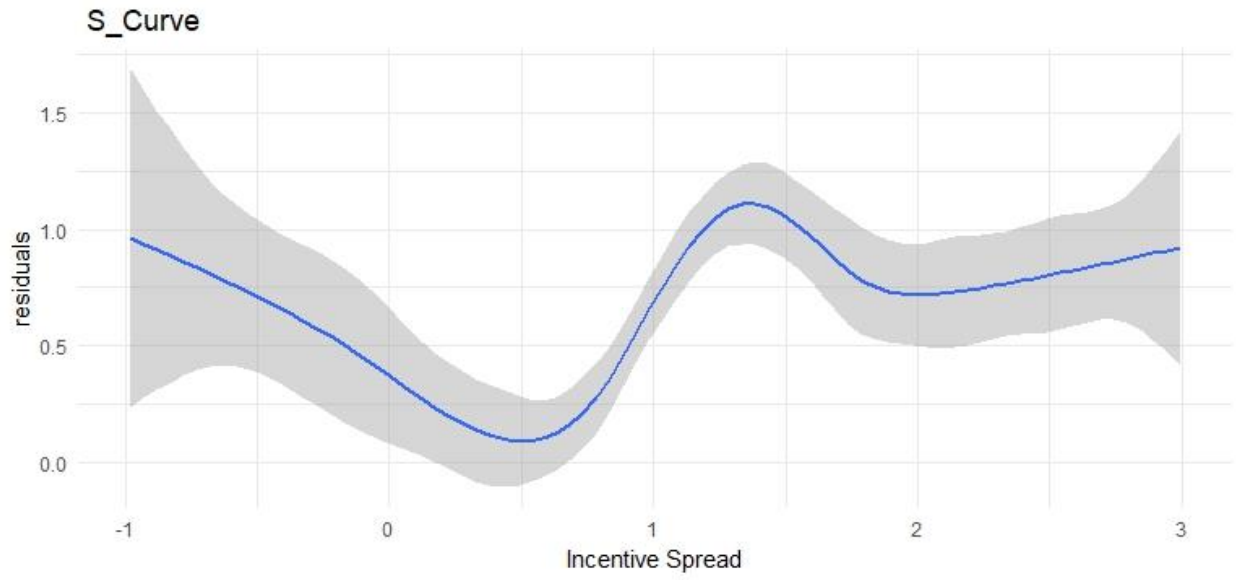
A. job promotion and capacity for higher monthly payment, leading to a more expensive home
B. moving to other places or losing employment
C. divorce and separation of borrowers
D. having more children resulting for the need of additional space



S-Curve

The importance of incentive spread is significant when we are looking at speed of response to positive economic opportunity. Generally, this response in graph is look like S-curve and sharper S-Curve shows faster prepayment rate. However, there are many ways to make the S-Curve, the raw response rate to incentive is what used in this research.

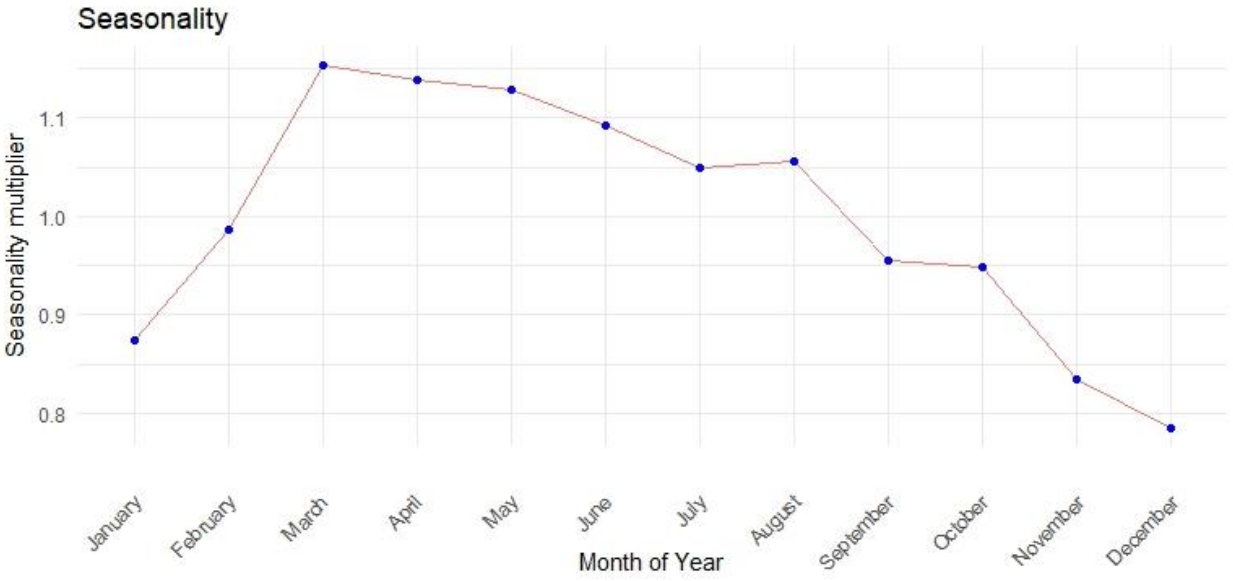
It means first, fit null model on responses by survival regression. The assumption is that incentive has normal distribution then analyses the deviance of residual according the incentive spread movement. For having more quality of shape, Drawing the plot, the smoother function is used for line and drawn in absence of outliers.



Seasonality

The other important risk factors in the model describe the seasonality associated with prepayments. There is a substantial difference between seasoning and seasonality. Seasoning is the change of prepayment behavior along the age of loan. Seasonality is about the periodic seasoning during each year, the seasonality depends on three main categories month of year, geographic location and economic condition.

It's well known that borrowers' propensity to move may vary from one month to the next. Therefore, in our model we consider the months of year of prepayment as a variable factor. Thus, we have twelve kinds of behavior. Instead of using cell-coding of these factors, I used proportion

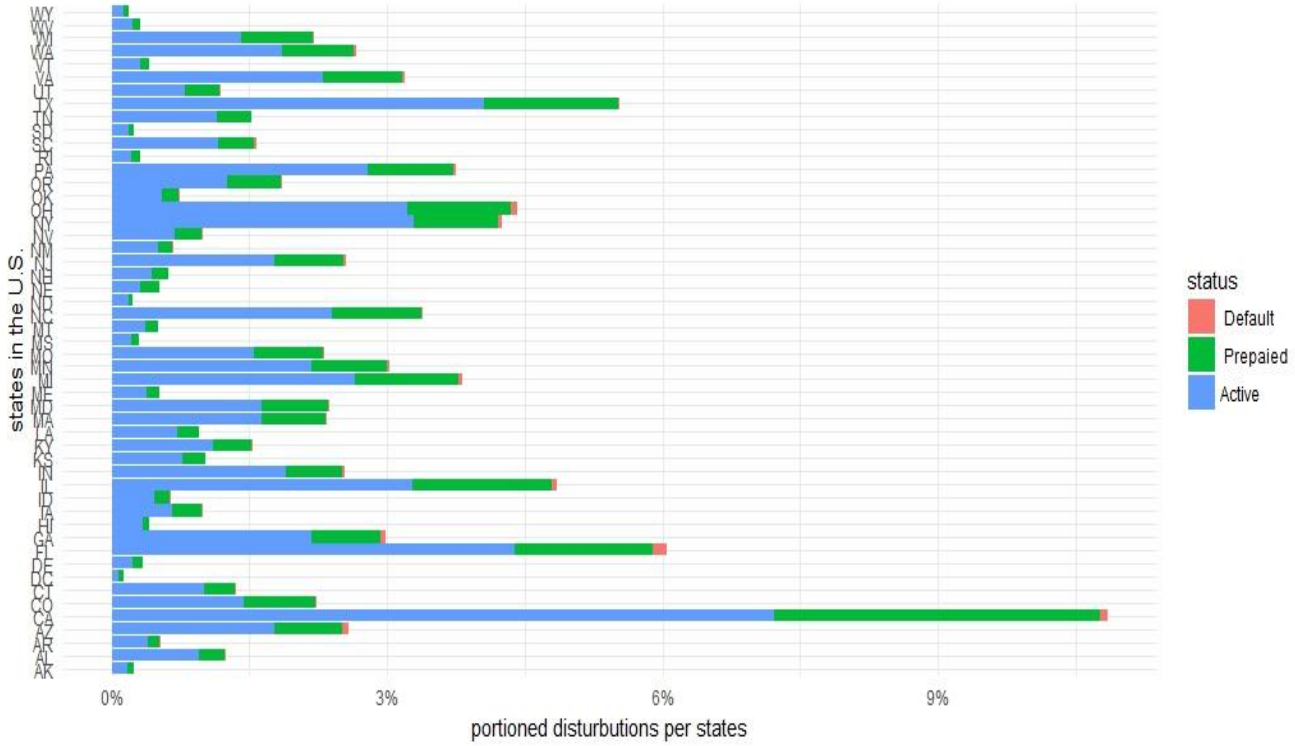


of prepayment in each month. A known famous example is school seasonality according to which in the summer, people tend to move more frequently.

Location (State of loan origination)

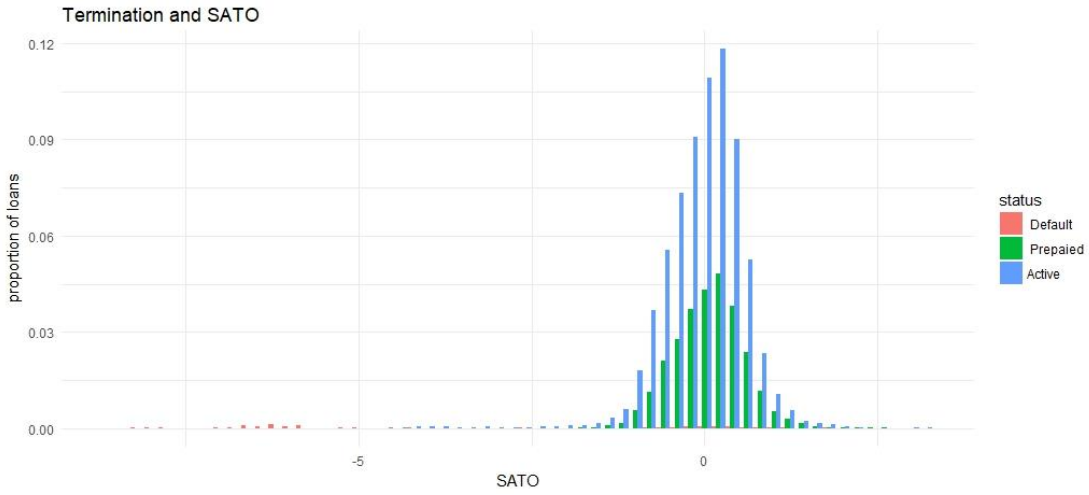
With respect to geographical location, in the dataset, the state of each loan has been provided. As is well known, each state in the United States can be treated using the same approach as the same way as months of years. In states with colder climate we expect different prepayment trends compared to hot climate states.

As mentioned in categorical treatments section for each loan *i* the probability of observing the loan in one of statues (contractual payment, prepaid, Default) is used to represent the variable in model.



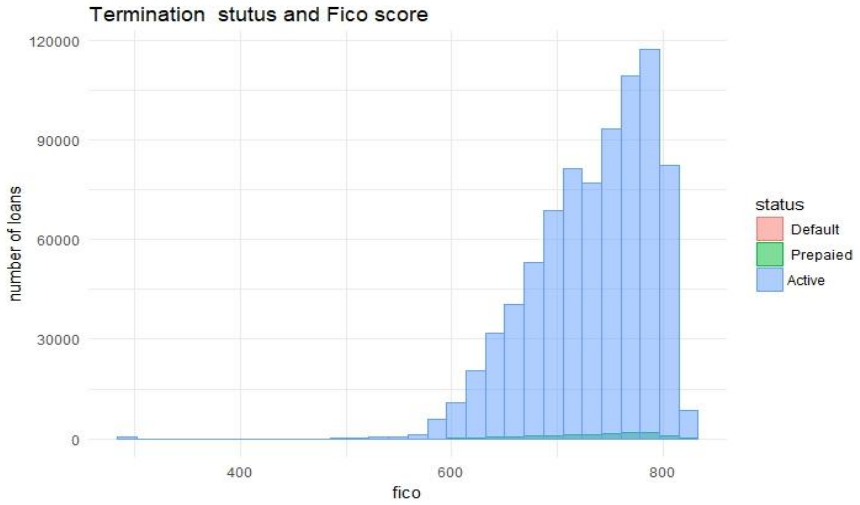
Spread-At-Origination (SATO)

To have a better understanding about the borrower's financial status, we need to look at spread between coupon rate and prevailing market rate at origination (same as financial incentive, but at origination). This variable can help us about the credit score cure effect. Most of borrowers at origination have lower age and credit score, but after 7th month payment they can have better coupon rate (without delinquency), with considering that in this data set Fico score and Debt to income are reported only at origination with adding this variable model can capture better analyses of borrower at origination comparing to mortgage market. Considering the plot for distribution of SATO, the minimum for the SATO is -3.97 and maximum is 3.37, while most of the loans a showing spread between -2 and 2. On the other hand most of loans outside of -2 and 2 did not experienced the event of termination. In result, loans with Sato spread of -2 and 2 has chosen that seems balanced and normally distributed. there is a similarity between SATO and financial incentive in case validating with response variable that more prepayments happen in positive side of the graph and residual plot versus parameter is S-Curve.



Fico Score

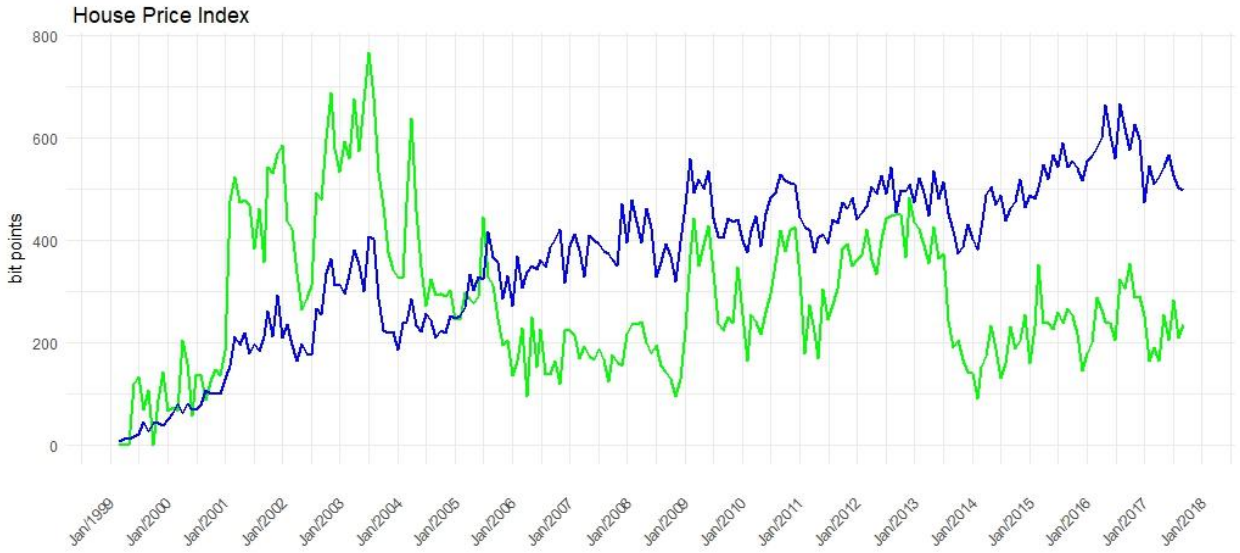
The Credit score is based on payment history of borrowers and some other factors, which capture the borrower's financial stability and credibility at the time of origination. According to Freddie Mac this score is not updated during the life of loan. In general, it is shown that, high



Fico borrowers tend to exit the pool faster because they likely have more options. Similarly, low score borrowers may also exit the pool quicker because of delinquency.
The plot of fico score for both is skew negatively. It seems those in last class of credit score (Fico) shown higher prepayment that others.
On the other hand, since the numerical values for this variable are significant, that cause to have very low coefficient in model that it is hard to analysis, so it scaled according mean of variable in modeling.

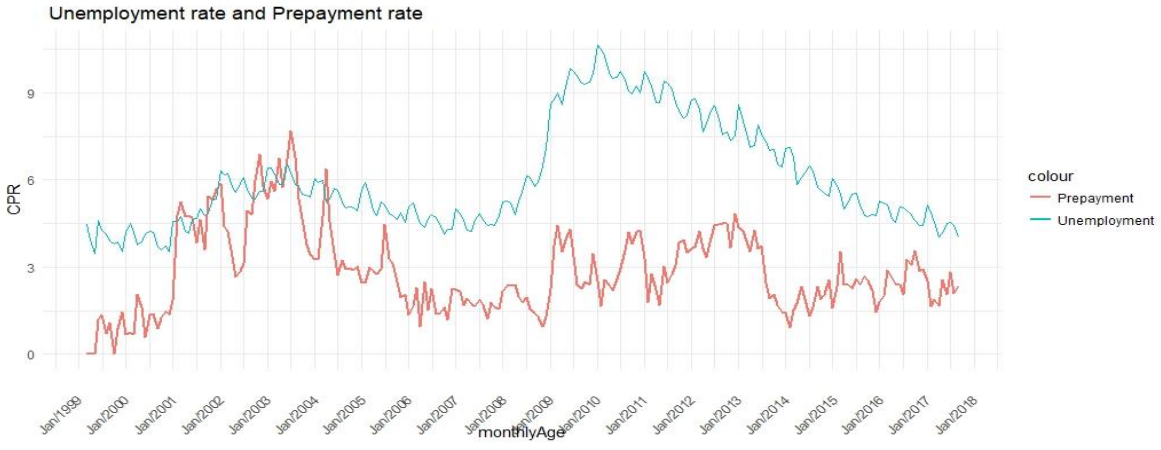
House Price Index (HPI)

House Price Index is a survey data that collected by government agencies to monitor the housing and economy. In case of prepayment is critical, since significant change in underlying house price can promote to prepayment. Cash-out if price rises and default if price falls.



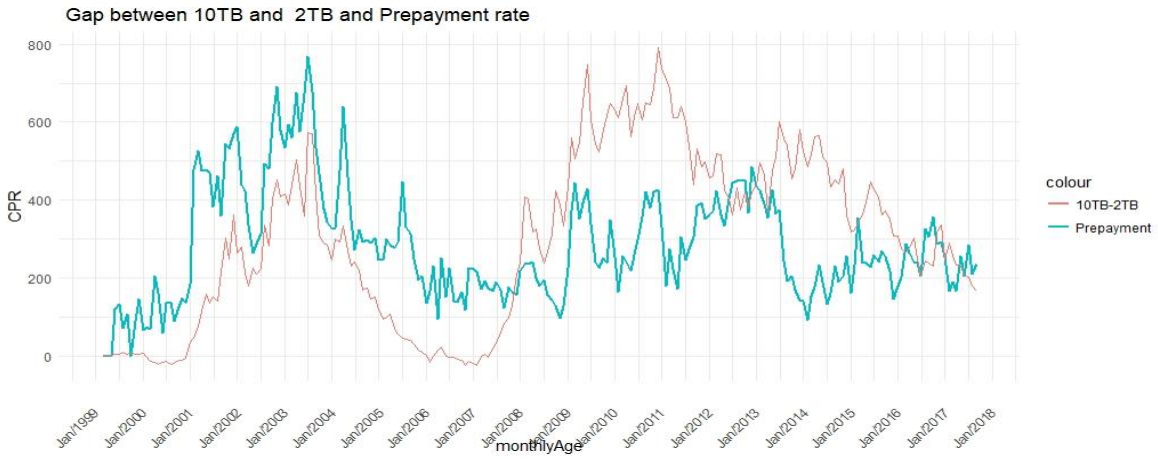
Unemployment Rate:

Unemployment rate is one of most important economic factor that influences in borrower's decision to stop the monthly payments. In our research we confirmed that higher prepayment rate and high unemployment rate happened in almost same time. In our study focus is on Single family borrowers and residential mortgage loans, therefore decision that lead to losing the house that family is living in is not easy even though if it is cause by losing the job. In plot below, average monthly unemployment rate and monthly prepayment rate shown. For each loan unemployment is been assign based on the state and calendar date of loan's performance.



Spread between 10-year Treasury Bill and 2-year Treasury Bill:

shows the market condition along the observation period. This factor shows the possible gap that is between long term investment and short-term investment.
Generally, 10-year Treasury Bill (10-TB) is considered Synchronized with 30-year fully mortgaged loan and 2-year Treasury Bill (2-TB) with 15yr fully mortgages loan of loan.
When spread between 10 TB and 2TB is wider, Financial there is more opportunity for loans with 30 and longer maturity to refinance to shorter loan with much lower coupon rate. The plot compares the spread and prepayment rate in our data, which provide visual proof to this relation.



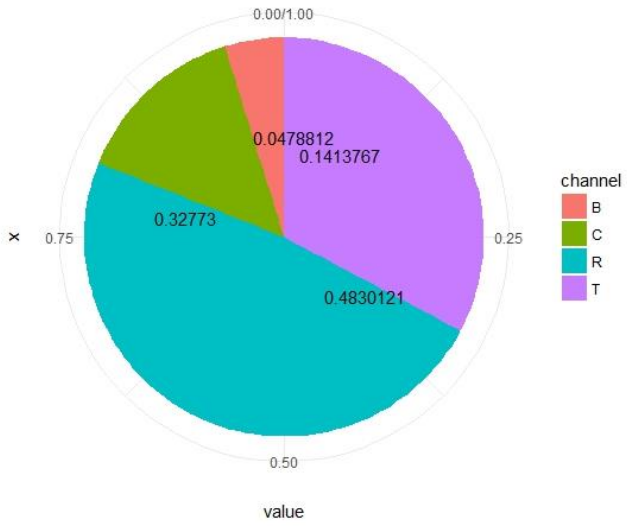
Channel : (**Disclosure Indication**)
Disclosure indicates whether a Broker or Correspondent, as those terms are defined below, originated or was involved in the origination of the mortgage loan.
Correspondent is an entity that typically sells the Mortgages it originates to other lenders, which are not Affiliates of that entity, under a specific commitment or as part of an ongoing relationship.
Retail Mortgage is a mortgage loan that is originated, underwritten and funded by a lender or its Affiliates. The mortgage loan is closed in the name of the lender or its Affiliate and if it is sold to Freddie Mac, it is sold by the lender or its Affiliate that originated it.

R: Retail

B: Broker

C: Correspondent

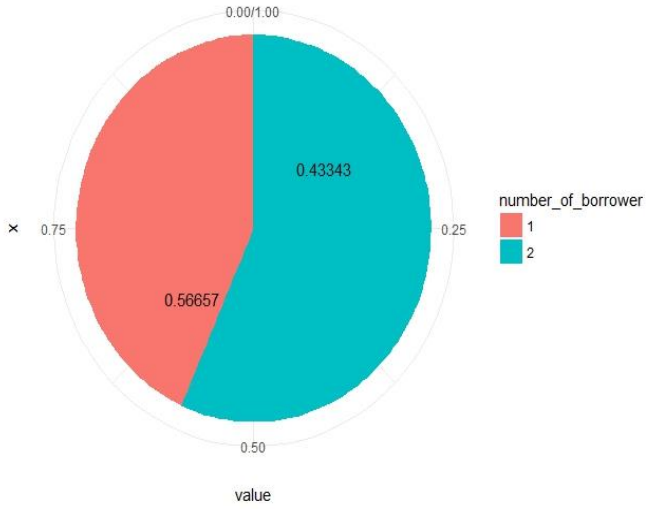
T: TPO (Third Party)



Number of Borrowers:

One of non- linear factors influencing in prepayment is Divorce rate.

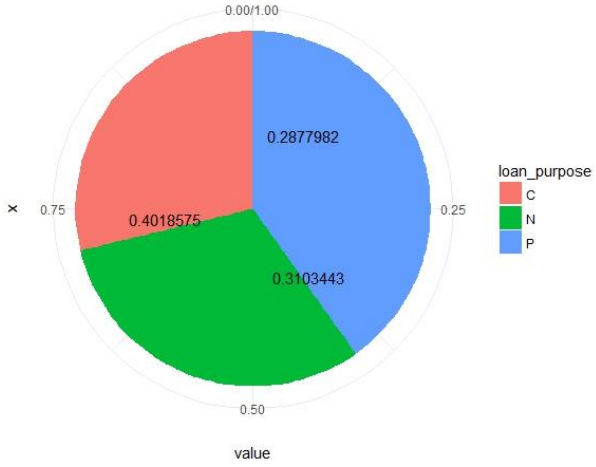
Since the focus in our data is single families, if a loan has two borrowers, it is assumed that borrowers are couple. Binary form of the variable can help us to consider the factor of divorce in the modeling prepayment.



Loan purpose:

The loan purchase suggests that the hazard ratios of the levels refinance and cash-out refinance are against the purchase level. Generally, a Cash-out Refinance mortgage loan is a mortgage loan in which the use of the loan amount is not limited to specific purposes. A mortgage loan placed on a property previously owned free and clear by the Borrower is always considered a Cash-out Refinance mortgage loan. Generally, a No Cash-out Refinance mortgage loan is a mortgage loan in which the loan amount is limited to the following uses:

Pay off the first mortgage, regardless of its age Pay off any junior liens secured by the mortgaged property, that were used in their entirety to acquire the subject property pay related closing costs, financing costs and prepaid items. As an exception to the above, for construction conversion mortgage loans and renovation mortgage loans, the amount of the interim construction financing secured by the mortgaged property is considered an amount used to pay off the first mortgage. Paying off unsecured liens or construction costs paid by the Borrower outside of the secured interim construction financing is considered cash out to the Borrower. (Freddie Mac, online documentation, 2018)



P: Purchase C: Cash-out

N: Not Cash-out Refinance
No Cash-out Refinance is limited to the following uses:

- Pay off the first mortgage,
- Pay off any junior liens
- secured by property.

Coefficients for Prepayment

VARIABLES	**values**	**p-values**	**Z-values**
INTERCEPT	19.08909	0.000000e+00	127.44125
INCENTIVE	-0.6698685	4.440892e-16	-8.104946
CNT_BORRODUM	0.4518749	5.434519e-06	4.547274
FLAG_FTHBDUM	5.442921	0.000000e+00	30.62549
DELQ-STSDUM	3.583689	0.000000e+00	27.22454
LOAN_PURPOSEN	0.04972473	0.712016663	0.369149
LOAN_PURPOSEC	0.1751045	0.1545108427	1.423778
LOAN_AGE < 12 MONTH	-0.7875807	2.532815e-05	-4.211856
12<LOAN_AGE <36 MONTH	4.5684804	0.03134822	16.257127
36<LOAN_AGE <72 MONTH	-10.64552	0.000000e+00	-74.88196
72<LOAN_AGE <195MONTH	10.066913	0.000000e+00	43.75975
LOAN_AGE <195 MONTH	9.671789	0.000000e+00	46.47597
CURRENT_LTV	-0.9404152	1.759478e-06	-4.779253
SATO	-0.5825207	1.28421e-06	-4.842182
FICO	-0.02556378	0.0000000000	-17.34440
CHANNAL C	4.501848	0.000000e+00	21.60441
CHANNEL B	0.9608718	6.115396e-04	3.426444
CHANNEL T	2.212546	0.000000e+00	20.61974
STATE	28.16114	0.000000e+00	3810.797
SEASONALITY	4.163325	0.000000e+00	8.705491
UNEMPLOYMENT	-0.005329858	0.8917154	-0.1361339
LTV	-7.015599e-03	0.000000e+00	-4.018633e+02
T10Y2YM	2.324647	0.000000e+00	19.53714

	Coefficients for Default		
VARIABLES	**values**	**p-values**	**Z-values**
INTERCEPT	12.84868	0.000000e+00	86.09203
INCENTIVE	-0.1750189	3.101455e-02	-2.156886
CNT_BORRODUM	0.1701610	8.477235e-02	1.723643
FLAG_FTHBDUM	5.279400	0.000000e+00	29.72288
DELQ-STSDUM	3.583689	0.000000e+00	76.40331
LOAN_PURPOSEN	0.04972473	0.007022736	2.695764
LOAN_PURPOSEC	0.1751045	0.0003057857	3.610349
LOAN_AGE < 12 MONTH	-3.9925166	0.000000e+00	-21.479970
12<LOAN_AGE <36 MONTH	0.6032829	0.03134822	2.152624
36<LOAN_AGE <72 MONTH	-11.49406	0.000000e+00	-81.66481
72<LOAN_AGE <195MONTH	2.639038	0.000000e+00	11.51192
LOAN_AGE <195 MONTH	11.385831	0.000000e+00	54.95493
CURRENT_LTV	-0.7428490	6.251184e-03	-2.734306
SATO	-1.0600222	0.000000e+00	-8.848124
FICO	-0.02629099	0.0000000000	-16.70209
CHANNAL C	5.157635	0.000000e+00	24.75599
CHANNEL B	1.1691193	6.115396e-04	4.185272
CHANNEL T	2.212546	0.000000e+00	24.75903
STATE	33.46354	0.000000e+00	4544.904
SEASONALITY	4.324354	0.000000e+00	9.095103
UNEMPLOYMENT	0.453897432	0.0000000	11.7971278
LTV	-7.015599e-03	0.000000e+00	-4.018633e+02
T10Y2YM	2.023072	0.000000e+00	17.15022

***Some Statistical Definitions*:**

Standard Errors (SE):
Standard Deviation / Square root of number of observation
Z-Value: model's coefficients/model's Standard Errors
P-Value: 2*(1-probability Normal distribution of absolute value of Z)

Variable's Table for the Prepayment and the Default and their statistical significance level in previous tables. Tables shows the intercept and coefficients for both cases, the Prepayment and the Default. We provide coefficients we can create the regressions for all states of loans.

$$Y_1 = \ln\left(\frac{\Pr(Y_i=\text{prepay})}{\Pr(Y_i=\text{contracual payment})}\right) = \alpha_1 + \beta_1 X_i \qquad \text{Intercept} = \alpha_1$$

$$Y_2 = \ln\left(\frac{\Pr(Y_i = \text{Default})}{\Pr(Y_i = \text{contracual payment})}\right) = \alpha_2 + \beta_2 X_i \qquad \text{Intercept} = \alpha_2$$

$$\frac{\Pr(Y_i = \text{prepay})}{\Pr(Y_i = \text{contracual payment})} = e^{Y_1} \quad , \quad \frac{\Pr(Y_i = \text{Default})}{\Pr(Y_i = \text{contracual payment})} = e^{Y_2}$$

$$\Pr(Y_i = \text{contracual payment}) + \Pr(Y_i = \text{Default}) + \Pr(Y_i = \text{prepay}) = 1$$

$$\frac{\Pr(Y_i = \text{prepay}) + \Pr(Y_i = \text{Default})}{\Pr(Y_i = \text{contracual payment})} = e^{Y_1} + e^{Y_2}$$

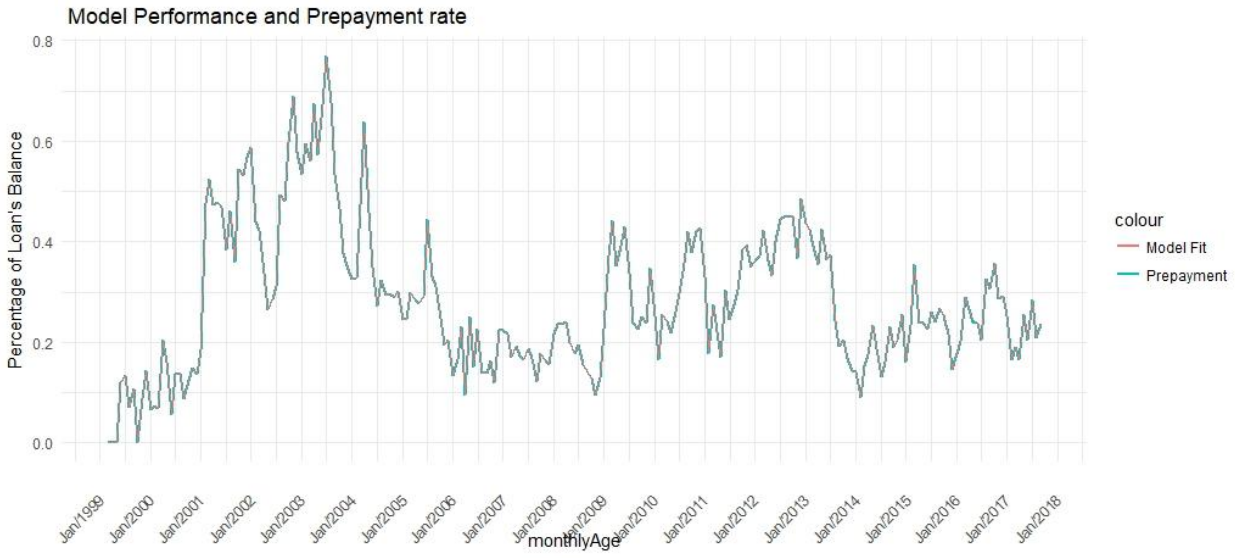
$$\frac{1 - \Pr(Y_i = \text{contracual payment})}{\Pr(Y_i = \text{contracual payment})} = e^{Y_1} + e^{Y_2}$$

$$\Pr(Y_i = \text{contracual payment}) = \frac{1}{1 + e^{Y_1} + e^{Y_2}}$$

Conclusion:

In this thesis multinomial logit model (MNL) is applied for estimating risk in a mortgage portfolio of Freddie Mac. , the multinomial logit model (MNL) is widely used in the prepayment literatures. The major disadvantage of the competing risk model is its limited ability to include time varying explanatory variables. The survival function has shown a relatively steep prepayment survival curve and indicated that the effect of loan age on prepayment rates is significant and not linear. To incorporate this, loan age is included in the MNL model as a categorical variable.

The option theoretic model can be constructed as a lookback put option on the mortgage. This is an interesting from a theoretical standpoint however its major drawback lies in the inability of incorporating the behavioural risk that resides in prepayment decisions. Such behavioural risk arises from the fact that mortgagees are unaware of optimal prepayment options as well as from diverging borrower specific attributes such as the Debt-to-Income ratio, FICO score, calender month seasonality, region in which the property is located and the divorce rate. In this research Exogenous factors are used to estimate prepayment rates according to a number of borrower specific, loan specific and macro economic variables. Including the put price (saving bucket incentives) as explanatory variable is a good alternative to incorporate the refinancing incentive however leads to less accurate predictions compared to using a straightforward variable for the refinancing incentive.



Selected Bibliography:

1. Economic Research Data of the Board of Governors of the Federal Reserve System, *https://www.federalreserve.gov/data/mortoutstand/current.html*, 20/07/2018
2. Frankj.Fabozzi, *The hand book of Mortgages-backed securities*, Oxford University, 7th ,2016
3. Glenn M.Schultz, CFA, *Investing in mortgage-backed and Asset-Backed Securities*,WILEY,2016
4. James Honaker, Gary King, *what to Do about Missing Vales in Time series and Cross-section Data*, American Journal of Political Science, Vol. 54, No. 2, April 2010, Pp. 561–581
5. Nicholas J. Horton and Ken P. Kleinman, *A comparison of missing data methods and software to fit incomplete data regression models, PMC 2007 Mar 30. final edited form as:Am Stat. 2007 Feb; 61(1): 79–90.*
6. P.D.Allison , *Measuring of fit for logistic regression*, SAS Global Forum, Paper 1485, 2014
7. Stanton, R. & N. Wallace (2011), *The bear's lair: Index credit default swaps and the subprime mortgage crisis*, Review of Financial Studies 24(10), 3250–3280.
8. Schafer JL. *Analysis of incomplete multivariate data. London*: Chapman & Hall/CRC; 1997
9. Terry M., Grambsch, *Patricia M, Modeling Survival Data: Extending the Cox Model*, springer,2000
10. The bond market Association (1999), *standard formula for the Analysis of Mortgage-Backed Security and other related Securities*,1999
11. Kolbe. Valuation of mortgage products with stochastic prepayment-intensity models. PhD thesis, Technical University of Munchen, Centre for Mathematics, 2008.
12. G. Rodriguez. Non-parametric estimation in survival models. Princeton, 2005
13. K. Gerardi and C. Hudson. Did nonrecourse mortgages cause the mortgage crisis? Real Estate Research Federal Reserve Bank of Atlanta, 2010.
14. Y. Deng. Mortgage termination: An empirical hazard model with stochastic term structure. Journal of Real Estate Finance and Economics, 14(3):309{331, 1997.
15. Y. Deng, J.M. Clapp, and X. An. Unobserved heterogeneity in models of competing mortgage termination. Social Science Research Network, 2005.
16. J. Green and J.B. Shoven. The effects of interest rates on mortgage prepayments. National Bureau of Economic Research, (1246):1{32, 1983.
17. M. Bissiri and R. Cogo. Modelling behavioral risk. Cassa Depositi e Prestiti S.p.a, 2014.
18. M. Consalvi and G. Scotto di Freca. Measuring prepayment risk: an application to uni-credit family financing. UniCredit Universities Working Paper Series, (05):1{35, 2010
19. E. Charlier and A. Van Bussel. Prepayment behavior of dutch mortgagors. Center Discussion Paper, Tilburg: Econometrics, 2001(64):1-33, 2001
20. J.P.A.M. Jacobs, R.H. Koning, and E. Sterken. Modelling prepayment risk. University of Groningen, 2005.

21. Gonchanov. An intensity-based approach for valuation of mortgage contracts subject to prepayment risk. PhD thesis, University of Illinois, Department of Mathematics, Statistics and Computer, Science, 2002.