Experiential Learning Project:
Leveraging Analytics to Improve Housing Stability in Hennepin County

Knowledge Transfer Documentation

April 27, 2018

Animesh Satyam, Bryce Quesnel, John Stephen A, Justin Hagstrom, Kevin Sorensen

CARLSON SCHOOL
OF MANAGEMENT

UNIVERSITY OF MINNESOTA

Contents

1 Overview

1.1 Description of Project

The purpose of this guide is to act as supportive documentation for the experiential learning project completed in Spring 2018 for Hennepin County as part of the Master of Science in Business Analytics (MSBA) program at the Carlson School of Management. With this guide we hope to pass on the methods, code, and insights that we were able to assemble over the course of the project so that Hennepin County can implement our process well into the future. The models contained herein were presented on May 2, 2018. Should you have further questions about this documentation, please contact:

- Bryce Quesnel or John Stephen A for technical questions regarding our predictive model or aspects of feature engineering.
- Animesh Satyam, Justin Hagstrom, or Kevin Sorensen for technical questions regarding our exploratory model and general insights.

1.2 Question

Hennepin County offers various emergency assistance programs to families and individuals at risk of being evicted from their homes. Unfortunately, the process for an individual or family to receive assistance takes approximately one month to process whereas an eviction can occur within two weeks of someone receiving notification of eviction. Additionally, evictions can end up being more expensive for the county than early housing emergency assistance intervention. It makes financial and ethical sense for the county to proactively anticipate evictions for the clients that they currently serve.

Given that eviction filings can happen much more quickly than the process to apply and be approved for emergency assistance, Hennepin County has engaged our team to answer the question: How can Hennepin County anticipate future county client evictions prior to an eviction filing so that the county can intervene by educating county clients about the emergency assistance programs available to county residents? We aim to allow the county to communicate about emergency programs with at-risk county clients with the goal of staving off potential eviction filings.

1.3 Data

We received approximately 25 tables each for "control" and "treatment" groups via text file from four separate databases – MAXIS, MMIS, SSIS Shelter System, and HMIS for data spanning 2008 through 2015. The treatment data was intended to be as complete of a dataset, split into 25 tables, for clients on a case which had received an eviction filing that the county could provide. The control data was intended to be a random sample of clients not impacted by an eviction filing. In this way, the county intended to provide us with labeled instances of eviction filings and non-eviction filings; however, we faced a handful of challenges with this approach by the county and these challenges may or may not limit the impact of our conclusions.

The first challenge with the dataset from the county does not own all data used in our analysis. Eviction filing court data is provided to the county from the county court system. When this data is provided to the county, the identifying information available from the courts is essentially limited to full name. For the county to identify which of its clients have experienced an eviction filing, the county must merge court record data to the county's data using the full name as the only matching information. Naturally there are many names that are duplicated in either system and which make merging data from the court's database to the county's databases problematic. If the county matched an eviction filing record from the court to more than one person in their databases, then the county did not send this eviction filing record to us for analysis – as the county would be very unsure which person in their data was the same person as the person in the court data. Another scenario could be that there are two people in the county with the same name and the first person is in the county's data. The second person with the same name could get evicted and the county would match the eviction filing record with the county's record for that name – even though the two people are entirely different. According to the county, over our time range 12,070 eviction filings out of 53,712 eviction filings from court records were matched with county records. Therefore, our analysis is based on only one quarter of the eviction filings in the county over our data's period and is limited to those instances in which one name in court records matched once name in the county's records. Consequently, our analysis is performed under the assumption that errors related to this matching method are ultimately few and inconsequential. This assumption does not eliminate potential selection bias, however.

Lastly, it is worth mentioning that many experts on eviction have identified landlord and address data as highly predictive data points. We did not have access to address or landlord data for our analysis, though the county expects to have this data available in the future. We expect the county to improve upon our analysis once the county is able to incorporate these data points into our models. Please see the Appendix for a list of the data tables provided by the county.

1.4 Approach

1.4.1 Predictive Modeling

We have used three predictive modeling techniques to predict, one month in advance, the likelihood of a case receiving an eviction filing in a given month. To achieve this goal, we have taken three separate approaches to provide models for low-, medium-, and high-cost interventions. At the outset of this project, Hennepin County expected to "intervene" in eviction filings by promoting its emergency assistance programs to clients imminently at-risk of an eviction filing. As communications take many forms and have many varied associated costs, each of our models can be deployed based on the expense of a preferred method of intervention. The low-cost intervention model prioritizes predicting as many at-risk households as possible with little regard for incorrectly predicting non-at-risk households as at-risk. Conversely, the high-cost intervention model aims to correctly predict at-risk households while simultaneously avoiding as many incorrect predictions as possible. For example, emails are relatively inexpensive to send, and the county will not incur many additional costs for reaching out to 1,000 clients versus 5,000 clients. If our high-cost intervention model identifies 1,000 clients to reach out to and only captures 10 at-risk clients, then our low-cost model might identify 5,000 clients of whom 20 are at-risk. If the intervention method is inexpensive (such as an email), then sending 5,000 communications and reaching 10 more at-risk clients could make sense. If the county has a high-cost intervention in mind

such as traditional mail or one-on-one communication, then an additional 4,000 communications to reach 10 more at-risk clients may not make financial or practical sense. Our models allow the county flexibility depending on the intervention method that the county is prepared to deploy.

All predictive models use time-varying features and require at least six months of data prior to the month in which the model is deployed. Time-varying features comprise many of the most predictive features that we were able to generate, although for many county clients we do not have enough data over time to include them in our predictive models. For this reason, we have created our second approach which consists of clustering our cases into different groups with varying risks of eviction filings without using time-varying features. This second approach is more valuable for gaining insight into general combinations of features that could factor into a higher likelihood of receiving an eviction filing.

1.4.2 Exploratory Modeling

Our second approach aimed at reducing the number of eviction filings in Hennepin County is based on clustering. We have taken any data available for cases and have considered demographic makeup of a household, average household income, average rental expenses over time, and other features and have created a dataset upon which to perform statistical techniques to segment our sample of cases into different clusters of varying propensity of receiving an eviction filing. By inspecting the segments with the greatest propensity of receiving an eviction filing, we have provided profiles of households that typically experience a greater risk of receiving an eviction filing. Our goal is to provide the county with knowledge of client characteristics that co-occur within at-risk populations. The county will be able to use these general insights to better target existing eviction prevention measures as well as any that may be developed in the future.

1.5 Technical Specifications

Our analysis is performed in R. Being an open-source tool, R is constantly changing as contributors expand its functionality and adapt the program to new and interesting problems. As such, our code was produced with R Studio having the following specifications. We recommend matching our R Studio specifications when first implementing our code.

R Studio Version: Version 1.1.383

R Version	
platform	x86_64-w64-mingw32
Arch	x86_64
OS	mingw32
System	x86_64, mingw32
Status	
Major	3
Minor	4.3
Year	2017
Month	11
Day	30

Svn Rev	73796
Language	R
Version.string	R version 3.4.3 (2017-11-30)
Nickname	Kite-Eating Tree

2 Data Preparation

2.1 Entity Relationship Diagram

Given the number of data tables provided by the county, we felt it useful to create an entity relationship diagram (ERD) to visualize the connections between the various data tables. We assume that the county is aware how each of the tables relates to one another, but we want to provide our ERD to eliminate any doubt regarding our approach to assembling our predictive and exploratory datasets. Please refer to the accompanying document, ERD.png for the full entity relationship diagram. The central table of the dataset is the Program Eligibility which contains programs for which each person/case is eligible in a given month. We joined every other table to this main table via a combination of Person ID (E_PMI), Case Number (E_CaseNumber), and month (BeginDate). Some data points are at the case level and do not vary among people on the same case. In these instances, we joined case level data to all persons using just Case Number and month.

2.2 Data Aggregation

2.2.1 Choice of Grain

As we approached aggregating the provided data into a single dataset, we first needed to decide the lowest grain, or level, of the data for our predictive and exploratory analyses. Due to the nature of the data, we could choose to aggregate our base dataset at the person, case, month (person) level or at the case month (case) level. The difference between the two is that data aggregated to the person level would allow analysis on every person in a household for every month in the data while aggregating data to the case level would require summarization of a household's attributes to one record in our data for every month. Since the latter method loses quite a bit of detail, our base file is aggregated at the person level. Additionally, most of the data was at the month level with two notable exceptions - shelter and emergency assistance program denial data which was provided at the day level. For the sake of simplicity and computational expense, we opted to aggregate all dates to the month level and instead counted the number of shelter days within a month and noted all months in which at least one emergency assistance program denial occurred.

By choosing the person level as our grain for our 'base file', we are still able to aggregate our data up to a less specific level (such as up to the case month level). We have the flexibility to look at people by year, or cases by month, or people/cases by quarter with simple aggregations from our base file. Though we produced a base file for initial exploration and for future aggregation flexibility, we did decide to perform our prediction at the case month level. This choice allowed us to predict eviction filings for whole households, rather than for individuals who may be part of a larger household. This choice makes logical sense and improves our model performance compared to predictions at the person level.

2.2.2 Building our Base File

As mentioned above we combined all the various data tables into one dataset for our predictive modeling and exploratory analysis. We refer to this combined dataset as our 'base file'. We are providing R Markdown files containing steps in R for creating a dataset similar to our base file. Please refer to the files titled 'R_Full_Merge_T.rmd' and 'R_Full_Merge_C.rmd', one being for the treatment data and one being for control data. Both files are relatively similar, the only difference being the treatment R Markdown includes data on evictions while the control file just fills those identical columns with 'NA' or blank values. By feeding each of these files the appropriate text documents, it will combine them and produce the base file that is to read in by our prediction.

Please see each individual file in the folder for specific details on how the code works. The code is commented in R Markdown format so each individual line's purpose is explained.

2.3 Candidate Features and Feature Engineering

Hennepin County has provided us with many features for use in our analysis. Below is a list of the types of features provided by the county as well as descriptions of the types of features that we have created from the original features provided by the county. Many of the most important features for use in the predictive models are features engineered from the data provided by the county. Following the discussion of the feature types below, we will have the necessary background to discuss our predictive models and exploratory approaches.

Categorical Features

Categorical features include predictors that were not represented on a numerical scale and were either factors or binary options. An example would be Non-English, which either is represented by either a 0 (if all members in the case speak English as their primary language) or a 1 (if at least 1 member in the case doesn't speak English as their primary language).

Numerical Features

Numerical features include predictors that are presented on an interval scale. An example would be CurrentRent which is the current rent for a case in the most recent month.

Cumulative Features

Cumulative features are features that count the repetition of an event or activity over time. An example would be months_on_HC which tallies the number of months on healthcare for each person on that case.

Change Features

Change features represent a change in some type of binary, categorical feature for a given case over a period of time, typically referring to program participation. When a case appears to begin participation in program, the case will get a +1 for the change feature related to that program. If a case leaves a program, then the case receives -1 for this change feature related to that program. If a case remains on (or off) a program with no change in participation for that program, then the change feature related to that program remains 0 for that case. We tested multiple time periods ranging from 3-12 months and found 6 months to be the period of time with the greatest contribution to predictive performance. An example of this type of feature would be Change_FS, which represents the change in Food Support

participation in the current month compared with 6 months prior. For the sake of modeling simplicity, in cases in which no data was available 6 months prior and we use a value of 0 since we are uncertain about the changes. Our clustering solution (see section 4.1) is an excellent alternative for finding people with a high likelihood of an eviction filing with data less than 6 months.

Ratio Features

Ratio features are like change features, except they are used to compare the value of a current feature with an aggregation (such as maximum value) for that feature's values over the entire period of the data. An example of a ratio feature is rent ratio, in which we take the current rent for a case and divide that by the largest rent value in the history of that case. If the ratio is near 1, then a client is paying near the most they've ever paid for rent. If it's near 0, then they're paying a much smaller portion of rent than they have previously.

3 Predictive Model Analysis

Note that while our productionized model and code is available in section 3.4 (Implementing the model), if you want to review specifics in this section with the test and control data we used, please use the 'modeling_code.rmd' file. This is also commented out for clarity but was built to run on the test/control dataset as opposed to a single real dataset like the production model is.

3.1 Algorithm Selection

Hennepin County requested many models for implementation with interventions of varying cost. As previously mentioned, we decided to provide Hennepin County with 3 models of varying performance – one for each a low-, medium-, and high-cost intervention. We had several considerations in mind when choosing appropriate algorithms. For the low-cost intervention model, we wanted to emphasize predicting many cases which would face an eviction filing while accepting a greater number of cases identified as being at-risk of an eviction filing even if not the case. On the other end of the spectrum our high-cost intervention model is meant to predict a high proportion of households as at-risk correctly and greater penalizes incorrect eviction filing predictions. We chose several algorithms to test and optimized the best performing ones for each a low-, medium-, and high-cost intervention. Below is a summary of the models that we tried and why we tried them:

Model	Advantages	Disadvantages
Logistic Regression	Simple to understand, computationally fast	Works poorly with lots of features
Naïve Bayes	Works well with large data, computationally fast	Sensitive to training data, high variance in results
Gradient Boosted Trees	Captures feature interactions well, reduces bias	Computationally greedy, runs slow, relatively black box
Random Forest	Results are interpretable, works well with lots of features, captures feature interaction well	Medium-slow to train, requires large amounts of data to work well
K Nearest Neighbors	Trains very quickly, easy to understand, easy to interpret	Handles poorly with a large number of features, performance was poor
Elastic Net	Runs fast, penalizes non-useful features, low variance	Data is not always linearly separable, so performance isn't as good as tree structures
Support Vector Machines	Can separate non-linear data well, handles lots of features well	Black box model, lots of parameter tuning, computationally expensive

From the models listed above, we ended up producing the best results with Naïve Bayes for low-cost interventions, Random Forest for medium-cost, and Elastic Net for high-cost interventions. Naïve Bayes was one of the few models that worked well with a high false positive rate, and it was relatively simple to train and use and therefore was perfect for our low-cost model since it captured the majority of eviction filings. Random forest provided our best overall result relative to kappa statistic (see section

3.3.4 for more details). The model provided valuable information on feature importance. It also had consistent performance and minimal variance, which is why we chose it as opposed to gradient boosted trees (which performed similar but had more variance across results). The high cost model we chose was Elastic Net, which essentially is a penalized version of logistic regression. It was effective because it does good feature selection and although not all of our evictions were linearly separable, it consistently performed at around 35% recall which we felt was an appropriate number for a high cost intervention.

Specific details about model performance are in the forthcoming sections of this document.

3.2 Sampling

We used samples of our data when training our models to maximize computational efficiency as well as to balance the true positives to false positives rate of our prediction. Because actual eviction filings are very rare on the case month level, we oversampled our data such that we ended up with a dataset containing cases with an eviction filing to cases without an eviction filing at a ratio of 1:2. In situations where the event of interest is as rare as an eviction filing, oversampling can lead to better results. In our case the models were able to better identify patterns that distinguish cases with eviction filings from those without. However, oversampling eviction filings to achieve a more balanced ratio of eviction filings to non-eviction filings could have a major impact on the false positive prediction rate. If you sample eviction filings evenly with non-filingss, you will predict many more eviction filings as the model will be trained to expect the occurrence of eviction filings more frequently. At the same time, if you train a model with a 1:3 ratio (eviction filings to non-filings) it will be less sensitive to eviction filings, resulting in fewer predicted eviction filings but also fewer false positives. We found the ratio of 1:2 (eviction filings to non-filings) produced the best balance for this prediction.

3.3 Model Validation

3.3.1 Training, Validation and Testing Datasets

Separating the data into training, validation and testing is very important for a prediction problem like the one we are working on. The goal of this is to make sure we don't overfit our model to predict very well with the set of data it's trained on but generalize very poorly when faced with new data. There are multiple ways to do this, but we focused on separating the data from a time perspective.

The provided data only included eviction filings in 2008 through 2015; all years prior and post did not have any eviction filing data so we chose to ignore them in our analysis. We also wanted to be sensitive to the years 2008-2012 since we noticed there was a higher rate of eviction filings during this time period (specifically 2008 and 2009), likely as a result of the nationwide recession during these years. To prevent our model "fitting" the data during these years too closely – and not generalizing to other years, we saved data for years 2014 and 2015 for data in which we could validate and test our models on, respectively.

2014 was saved for validation, meaning we use that data to test the generalizability of our models produced on data from 2008-2013 to a new year not seen by the algorithm run on the training set. Essentially this is how we evaluate and make sure our models aren't under-generalizing to new data.

This is extremely important as all predictions in practice will be performed on data that the predictive model has not yet encountered. Additionally, we have a test set (2015) which represents our 'final results' and is only used for very special occasions to determine results (midterm presentation and final presentation). This is meant to represent how our model would run on an entirely new set of data that was not used at all during the building/validation process.

The timeline below visualizes this. Note that grey represents training data, green is validation data and blue is testing data.

3.3.2 Performance Evaluation

Due to random sampling in our oversampling method described above, the results below may vary slightly when reproduced by the county, although general trends should be relatively stable when the model is re-run. We tested our models' performance with both validation and test sets (as described above) to make sure our results were not simply overfit to our training data and that our models were generalizable over time.

As we were focusing on multiple models and on providing a variety of models for different intervention costs, we relied on confusion matrices to visualize the tradeoffs across many different models. Confusion matrices allowed us to view true positives, true negatives, false positives, and false negatives all in a single view. Beyond providing those, we also wanted to use both precision and recall which are measurements that indicate how well a model can make predictions. Recall represents the fraction of eviction filings predicted out of the total number of eviction filings occurred (true positive / (true positive + false negative). Precision represents how many true eviction filings we captured divided by the number of eviction filings that we predicted (true positive / (true positive + false positive).

Naturally the county would like to correctly predict as many eviction filings as possible while minimizing the number of incorrect predictions. Therefore, we felt the kappa statistic was the best overall measure of performance for our models as it balanced our predicted accuracy well against random accuracy.

$$kappa = \frac{totalAccuracy - randomAccuracy}{1 - randomAccuracy}$$

Our models' performance based on kappa statistic is below. As shown in the table, our Random Forest model had the best kappa statistic, although Elastic Net was only slightly lower performing. We felt both models did a similar job overall, and either would work well depending on the cost of the provided intervention. The Naïve Bayes had a much poorer kappa statistic due to its high false positive rate. We do not recommend using this model unless the cost of intervention is very low.

	Intervention Cost	Total Eviction Filings	Eviction Filings Identified	Eviction Filings Missed	Annual Communications Provided	Kappa Statistic
Naïve Bayes	Low	396	**276**	**42**	84,414	0.0018
Random Forest	Medium	396	**159**	**159**	18,003	0.0126
Elastic Net	High	396	**110**	**208**	12,463	0.0125

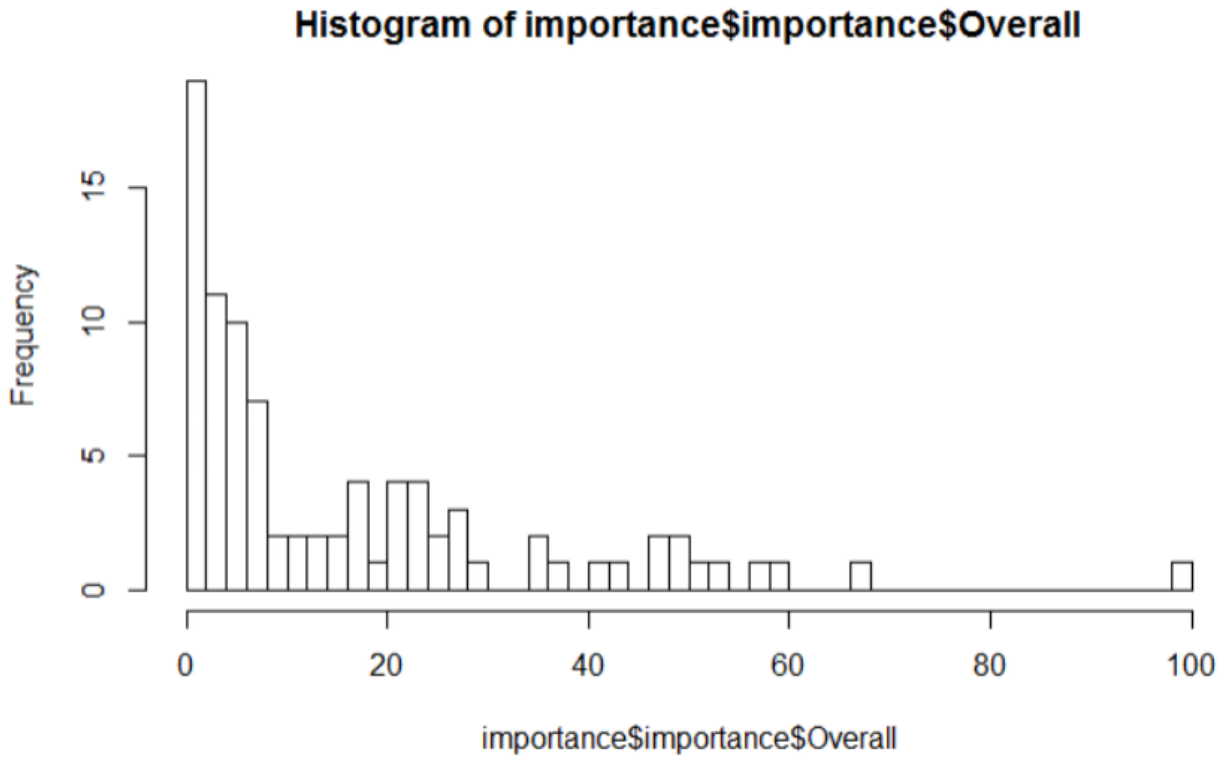
3.3.3 Feature Interpretation

We determined important features in two key ways. First, we used a random forest to understand which features provided the most value in the prediction of eviction filings and which were not useful. We felt a random forest was a good statistical technique to use because it was repetitive in nature (so it tested feature importance many times across many subsets of data) but also it produced a high performance result which was not the case with more traditional models like a logistic regression. Additionally, it does not run into constraints around heavily correlated variables that a traditional regression technique does; this is important because many of the features related to eviction filings are highly correlated (income, household size, etc.) but provide different levels of information and should be included in our prediction.

We used the Gini coefficient to represent variable importance, which measures how eviction filings are distributed amongst variables. In the case of a random forest, importance is recognized as the 'actual decrease in node impurity is summed and averaged across all trees' (Lee, Ceshine). Essentially, it represents how important that feature was for splitting eviction filings away from non-filings across all the 'trees' in the random forest; a higher number means more important (on a scale of 0 to 100). If you want to understand further, the article linked in reference written by Ceshine Lee goes into other importance metrics and even talks about the specific random forest package we used in our prediction (ranger).

For those features that were deemed important in the random forest, we then ran through a logistic regression to understand how they influenced eviction filing (either increasing or decreasing likelihood). Note that there is still some correlation present amongst the features; however, since we are just looking for a direction of influence and not a magnitude this should not interfere with our results. Additionally, this analysis is performed on randomly sampled data, so we expect minor variation in repeated analysis (though consistent results around significance and coefficient direction). Therefore, although the coefficients may not be entirely accurate in magnitude, so we don't recommend these models for high performance prediction.

We have included both significant and non-significant features to show all attempted methods. Note that non-significant features will not have a coefficient. One additional item to note is that for the sake of computational power, we calculated feature importance using random sampling as we discussed in section 4.2.2.



The graph above represents the distribution of 'importance' by features. Note that there are some very important features but 37 of the features have an importance lower than 5. We disregarded features with less than 5 importance from our logistic regressions to remove a lot of the correlation for better interpretation of the direction of coefficients on the 52 remaining important features.

Below we have the top 12 features ranked by importance. You may review all 91 tested features by referring to the Appendix for a more complete table.

Base Data	Feature Definition	Coefficient Direction	Importance	Significance	Type
Current_Rent	Current max Rent Expense for a case in that month	increase	100	Important	Numerical
current_avg_age	Calculation of the mean age for all people currently on the Case	Increase	67.1304905	Important	Categorical

Months_on_FS	Cumulative months on Food Support for that case up to that date	Decrease	58.4735736	Important	Cumulative
non_english	Indicator if there is a non-english speaker present in the case at the current time	Decrease	56.8324854	Important	Categorical
Previous_denials	Cumulative previous denials for either emergency assistance or emergency general assistance over the entire data period	Increase	52.2656257	Important	Change
Months_on_HC	Cumulative months on Health Care for that case up to that date	Decrease	51.924112	Important	Cumulative
cur_months_on_HC	Indicator of if the case currently on a Health Care program	Increase	49.7663029	Important	Categorical
curent_case_benefits	Current case benefits for a case in that month	Decrease	49.4946953	Important	Numerical
ProgramPayment_Ratio	Current Program Payments / Max Program Payments for that case across all time	Decrease	47.9137042	Important	Ratio
RentRatio	Current Max Rent In Household / maximum rent for that case across all time	Increase	46.5050135	Important	Ratio
cur_months_on_FS	Indicator of if the case currently on a Food Support program	Increase	43.3777181	Important	Categorical
cur_months_on_MF	Indicator of if the case currently on a MFIP program	Increase	40.292171	Important	Categorical

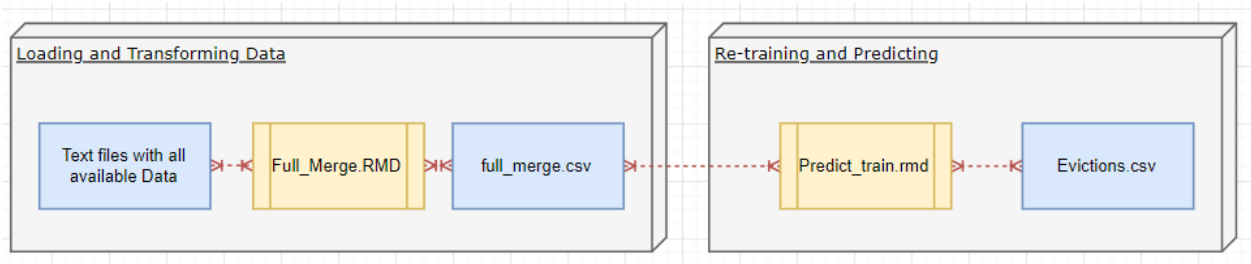
3.3.4 Feature Usage

For our Random Forest model, we used all features possible. Random forests split nodes based on the best available feature from a given subset of data (sampled with replacement from the entire training set), therefore non-useful features will not be used unless they are valuable in that subset of the data.

For our Elastic Net model, we implemented automatic feature selection. We wrote code to select features with a Gini coefficient importance greater than 5, and therefore if the model is fed vastly different data upon implementation at the county, then a few of the features may change slightly. Our version used only features deemed important in section 3.3.6.

For our Naïve Bayes model, we used all features as Naïve Bayes typically computes quickly with large amounts of data.

3.4 Implementing the model



Above is a visual representation of running the model. We tried to make the process as streamlined and simple as possible. To run a prediction use the following:

1. Full_merge.RMD R Markdown file
2. Predict_Train.rmd R Markdown file
3. .Txt files for all original data sources

A brief summary of how to run the model to predict eviction filings is present in the file 'Full_Merge.rmd'. In this R Markdown file, the user only needs to specify the .txt documents that contain the data to be predicted on and select the desired model type and it will create a CSV file full of predictions. This is the file that the prediction code will use.

From there, open the 'predict_train.rmd' R Markdown file and specify dates to train the model for (ideally as much time as possible) and date to predict for (the current month for example). After setting those limits, run the remaining code blocks to prep the model and build all the features, then select one of 3 final code blocks (high-, medium- or low-cost intervention model) to produce a CSV with eviction filing predictions.

3.4.1 Instructions for using the production model

The model is set to predict on all cases within the latest month in the dataset. For example, if you add in data from 2008 to 2016 it will output a file with eviction filings predicted for everyone present in the data in December of 2016.

Note that it is important to include as much data as possible for the models to train; abbreviating the data period to only a year or two of data will have a major impact on prediction. Please include data for as many years as possible when training models, while still retaining significant amounts of data for validation and testing. Each individual R Markdown file could take anywhere from 1-4 hours to run depending on the size of the dataset used. With our sample data from 2008-2016, it typically took around 2 hours to run the model. As the dataset size increases, so will the time to run the model.

Most of the prediction process is automated. All you need to do is run each code block by clicking the green arrow in the top right for that R Markdown. There are three things that need to be done in the code in order to predict eviction filings.

3.4.1.1 Change file names to load data

This occurs in the full_merge.rmd file in the second chunk. Follow the instructions in the code to complete this step. This step should include all data with no separation for test/control like was done in the original project.

One thing to note is that the assumption is that the county will format data for use in our models using the same format as our 'test' dataset referenced earlier in this document. We also expect the county to use up-to-date court data. If there are portions of the data that are not up to date this will result in 2 things:

1. The model should not be trained with months that only have partial data, and some with full data. If that occurs the model will not train properly as it will include the lack of information in certain months as a telling trait of eviction filing/non-filing. Therefore, make sure all data is present for all months (with the exclusion being eviction filing data in the most recent month)
2. If you are missing court/eviction filing data and train on months without it, there will be a bias for non-filings. Be sure to exclude any months in training in which there is not data available.

3.4.1.2 Select your train date range and prediction dates

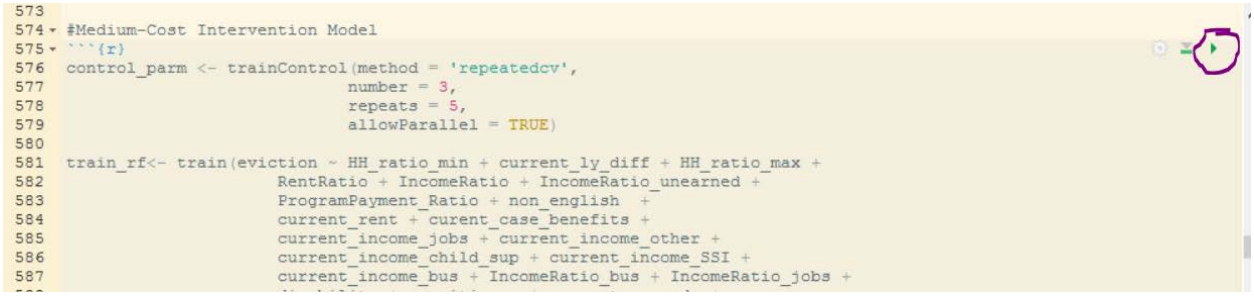
```
#Selecting Train/Predict Period
```{r}
#Enter current period for prediction
predict_date <- '01/01/2018'

#Enter start of training range
start_train <- '01/08/2018'
end_train <- '12/01/2016

```
```

The second code block in predict_train.rmd requires the user to select the training and prediction periods. Note that the training range should be as large as possible as variables perform best when they can take long periods of context into account. Additionally, as this model is meant to run at the monthly level, predict_month should be the period in which you want to predict in the specified format (in the mm/01/yyyy format, with day being the first of the month).

3.4.1.3 Select model you want to use (high, medium, low)



```
573
574 #Medium-Cost Intervention Model
575 ```{r}
576 control_parm <- trainControl(method = 'repeatedcv',
577                              number = 3,
578                              repeats = 5,
579                              allowParallel = TRUE)
580
581 train_rf<- train(eviction ~ HH_ratio_min + current_ly_diff + HH_ratio_max +
582                      RentRatio + IncomeRatio + IncomeRatio_unearned +
583                      ProgramPayment_Ratio + non_english  +
584                      current_rent + curent_case_benefits +
585                      current_income_jobs + current_income_other +
586                      current_income_child_sup + current_income_SSI +
587                      current_income_bus + IncomeRatio_bus + IncomeRatio_jobs +
```

The final 3 code blocks in the predict_train.rmd are the different cost models. Each one includes notes on what its main trade-offs are, but after running all code blocks up to those models, running a specific code block will output a .csv with binary outputs for which case are/aren't predicted to be evicted.

4 Data Exploration

4.1 Clustering

Within our data, many cases have continuous data for program participation over many consecutive months up to a month in which an eviction filing occurred. For example, a county client could have participated in the county's food support program continuously for many months and then eventually had an eviction filing against her/him without any break in program participation. This type of continuous data over time for a client is ideal and we can use such client profiles in our predictive algorithms. Not every case for clients is so consistent though. Often a client will be in the county's data for many months before leaving all programs. Sometimes such former clients then experience an eviction filing at some point in the future. For these clients, the county does not very accurate data at the time of eviction filing. For these types of clients, predicting eviction filings is not feasible as the county would not have current data at the time of such a prediction being created.

In fact, there were many cases that belonged to this category where prediction was not feasible as the cases did not have a significant number of months of continuous data for creation of the time-varying attributes used in prediction. We still want to provide a way for the county to help reduce eviction filings for these people. Therefore, we are presenting a clustering solution that allows the county to identify features of cases that co-occur and are associated with cases that have experienced a high rate of eviction filing in our dataset. Our hope is that the county can use general insights from our clustering solution to reach out even to non-county clients who meet our at-risk profiles. However, our clustering solution is not perfect and is meant as a general appraisal of features correlated with a higher risk of eviction filing – within the dataset that we have been provided. The county should use these general insights as a starting point for further research.

Data Used

Case level data with all the demographics, county programs, income, expense and case benefit information was used.

Attributes Used in Clustering	**Definition**
Case Family Size	The count of individuals on a single case
Highest Educated	The highest education level of a person in a case
Disability Ratio	The ratio of disabled people in a case
Average Instances of Program Count on Case Level	Count of all months of each program divided by the number of people in the case
Gender Ratio	The ratio of male to female person in a case
Average Age	The average age of all the people on a case
Average Utility Expenses	The average rent and utility expense for a case
Average Income (Business, Jobs, Unearned)	The average income from various sources at the case level
Average Case Benefit Payment	The average amount of case benefit payment received on case level
US Citizen Ratio	The ratio of people in a case that are US Citizens

English Speaking Ratio	The ratio of people in a case who speaks English

Clusters

We used k-Means algorithm to find the clusters of the cases. Any categorical attributes were converted into factors and all numerical attributes were converted to a common scale so that no single feature overpowered any other.

The team found 18 clusters. Four of these clusters had cases with eviction filings in various proportions while the rest 13 clusters had minimal or no eviction filing cases.

Below is an overview of the clusters with eviction filings and non-filings:

Cluster 1 (67% eviction filing): This group has the highest risk of an eviction filing with 67% cases having experienced an eviction filing at some point over during our data's time range. The group contains mostly single adults in their 50's who have some amount of high school education. More than 50% of these people have no income from jobs, unearned income, or business income. The average household income is around 300 dollars per month. These cases also have average total monthly case benefit payments of 150 dollars. Additionally, most of the people in this group are disabled as flagged in the data.

Cluster 2 (56% eviction filing): This group contains cases of which 56% have experienced an eviction filing over the course of our data. The group mostly contains small, young families with 3-4 person per household and typically the highest education level in the family is high school graduation. The average job-related income for the household is 350 dollars. There is no considerable income from business or other sources. The average rent for the group is 600 dollars per month with same amount of average case benefit payment. This group has very high interaction with healthcare(HC) and MFIP programs.

Cluster 3 (25% eviction filing): This group has a relatively lower number of cases having experienced eviction filing compared to the two previous groups. The group is mostly comprised of families with two members. Interestingly, the overall proportion of female to males over all cases is quite high at around 70% female. We suspect that this group includes some couples but mainly single mothers with one child. The average household age is 30 years and the highest education level in the family is completion of high school. The average household job income is relatively high for this group at approximately 1000 dollars per month but there is quite a lot of variability from case to case within this cluster. The average rent for the group is around 500 dollars per month and the average total case benefit payment for the household is 150 dollars per month. This group has very high interaction with healthcare (HC) program.

Cluster 4 (15% eviction filing): This group has the lowest eviction filing cases of our four clusters with eviction filings. This group consists of mostly male singles with an average age of 35 years. There is a high proportion of no income people. The average rent is around 350 dollars per month with average total case benefit payment of 150 dollars per month.

The clusters that we found that contained cases with essentially no eviction filings are high in number (13 groups). Rather than discuss each of them individually, some high level characteristics of these groups are below:

- Many non-US citizens whose primary language is something other than English
- Families of two are well represented – potentially couples
- Average household income of around 1100 dollars per month

- Otherwise we typically see larger families with high job-related income and high program interaction face much lower rates of eviction filings

Using the clusters:

There are two ways in which the cluster can be used:

Static: The above defined group characteristics can be used to put people into clusters and mark them for their risk factor. Even if their income or other information may not be available for a span of time, they can be characterized on other attributes and just an idea of their monthly income can tell us how likely is that person to be evicted in future.

This way the general awareness can also be increased with use of billboards, pamphlets or other such measures. If an area has higher proportion of people with characteristics close to high risk groups they can be made aware using this method.

Dynamic: The second method is to integrate the new incoming household or individual in to the system with data available at hand and then use the procedure below to put them into a cluster/grouping:

1. Once the data for the individual or the family is in the treatment/control group, run the full merge code(Full Merge.RMD) to get the complete data in the required format.

2. Once the full merge file is in place, run the segmentation code(Segmentation.RMD) to get the new clusters with the added data.

3. Once the code is executed, the cluster output file can be checked to see which group/cluster the individual or family belongs to.

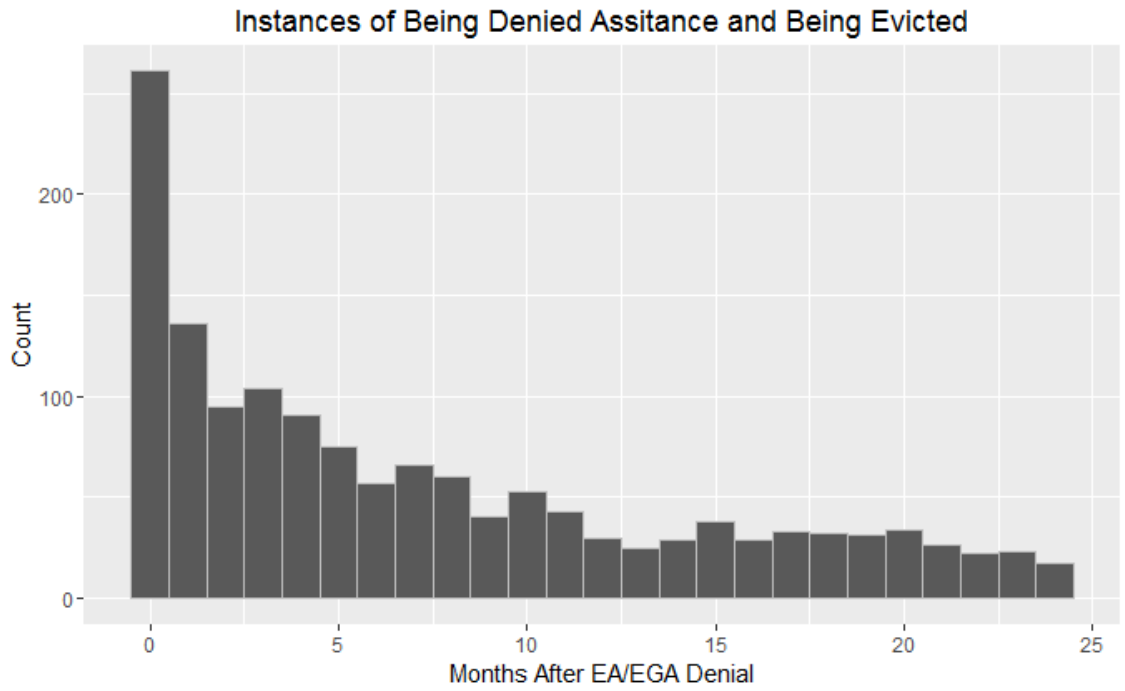
4.2 Intervention Awareness

As discussed earlier in this document, Hennepin County approached our team with the idea that we would be producing a predictive model to predict the likelihood of eviction filings for county clients. The general idea is to predict the likelihood of eviction filings to target communications to county clients at risk of facing an eviction filing. Imbedded in this idea is the assumption that many Hennepin County clients who face an eviction filing do not know about emergency assistance prior to an eviction filing. Otherwise it would not make sense to reach out to these households and inform them of county emergency programs.

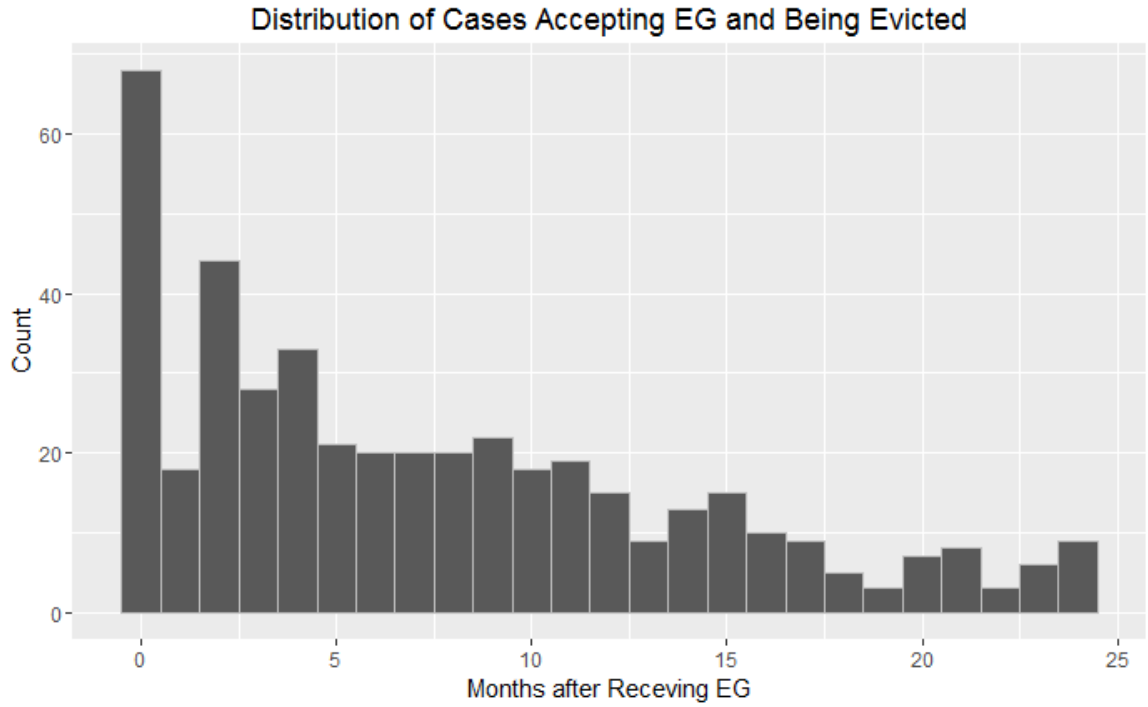
As part of our analysis (see the Eviction Summary Stats.RMD file), we looked deeper into the cases with eviction filing to see if these cases interacted with either emergency assistance (EA) or emergency general assistance (EGA) prior to facing an eviction filing. Upon inspection we found that **81% of cases applied for either emergency assistance or emergency general assistance prior to facing an eviction filing.** Therefore, spreading awareness of emergency programs to county clients may not be an effective intervention for preventing eviction filings.

The following graph is of households that were denied either EA or EGA and later went on to have an eviction filing within 24 months. It shows the number of months between their last assistance denial and an eviction filing.

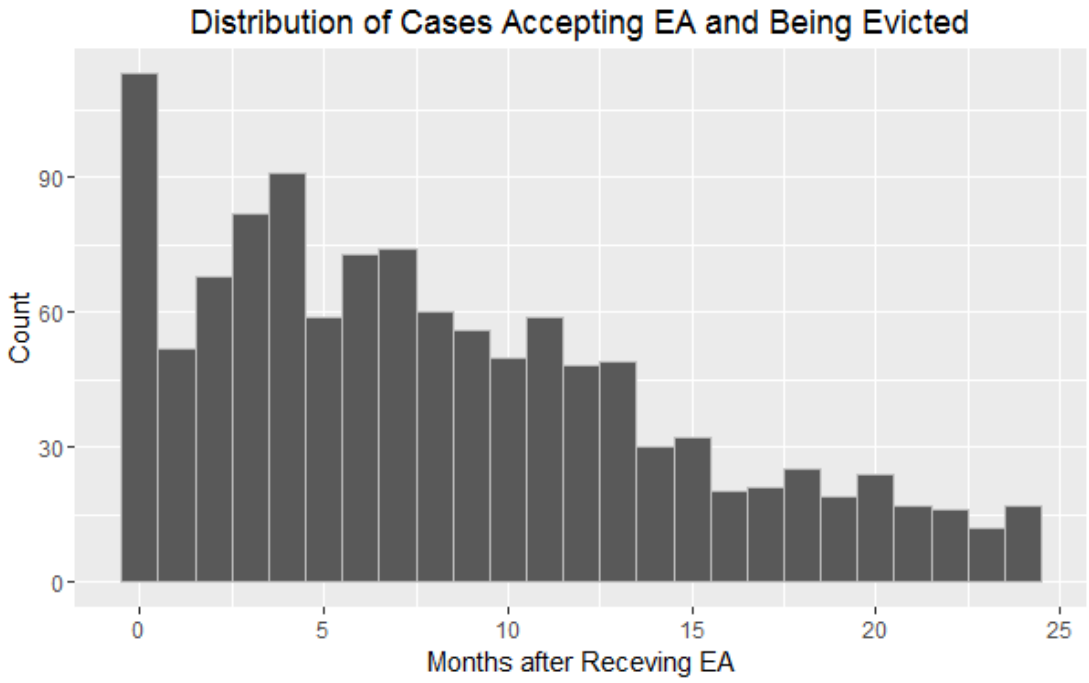
1450 total instances of a case being denied assistance and going on to be evicted.



The following is similar to the graph above; however, it is for cases that accept EG and went on to be evicted. 443 total instances of cases receiving EG and going on to be evicted.



This graph below is like the two prior graphs. There are 1523 instances of a case accepting EA and going on to be evicted.



Among these three graphs, there is some overlap. I.E. there are not 1450+1523+443 distinct instances that have interacted with the emergency assistance programs. When accounted for distinct cases however, there are 2588 distinct instances of a case interacting with the programs and going on to be evicted.

5 Future Steps

5.1 Refine models to include landlord and address data

The research literature on evictions informs us that certain locations of a city are more likely to have high eviction rates due to the low socio-economic conditions endemic to these regions. Another important factor that plays into high eviction rates is the influence of landlords over their tenants. Research shows that some landlords are more likely to evict tenants to hoard deposit money. Data points capturing this type of predatory landlord behavior along with zip code-specific details would play a pivotal role in improving accuracy of the model. Therefore, the county should supplement our analyses with address and landlord data.

5.2 Reduce eviction filings through other strategies

We have concluded that spreading awareness about EA/EGA programs to county clients is not the most effective means of preventing eviction filings. Most cases facing an eviction filing are already aware of the existence of these programs. Given this, the county should develop alternative ways of reaching out to potential households at risk for an eviction filing. The county could re-emphasize the role that caseworkers play in high-risk county clients that come under their purview. Alternatively, the county could investigate the effectiveness of emergency assistance programs in general.

6 References

Lee, CeShine. “Feature Importance Measures for Tree Models - Part I.” Medium, The Artificial Impostor, 28 Oct. 2017, medium.com/the-artificial-impostor/feature-importance-measures-for-tree-models-part-i-47f187c1a2c3.

7 Appendix

7.1 List of Data Tables from Hennepin County

MasterEvictionsList	List of every eviction in Hennepin County
MatchSummary	Name Match Summary for the 4th judicial court for Hennepin County eviction case defendants and distinct PMI count for those matched. This includes the basic process followed for the name matching.
Court	List of evictions court case data for a case defendant; includes Current_Case_Number (encrypted), Case_Filed_Date, Party_Full_Name (encrypted), Judgment_Date, Judgment_Type_Code, Judgment_Type_Description. If data was matched to the eligibility system, this includes matched type indicator, and de-identified person identifier (E_PMI) if matched for the eviction case defendant. Data is restricted to evictions cases identified as associated with the 4th judicial court for Hennepin county.
Matched	List of de-identified persons (E_PMI) with exactly one matched record against court data with matching based on name; includes person characteristics [non-time variant] such as Gender, BirthYear, DeathYear, Ethnicity, Race, Language, HasNonHealthCareProgramEligibility, and HasHealthCareProgramEligibility.
EADenials	List of all Emergency Assistance program application denials for those matched persons on associated cases.
PersonEligible	List of eligible persons (E_PMI): eligible evicted and eligible non-evicted on the same eligible case; includes person characteristics [non-time variant] such as Gender, BirthYear, Ethnicity, Race, Language, HasHealthCareProgramEligibility. This list was built by identifying all the eligible cases the matched person (eligible evicted) was eligible under for various programs for time period of interest, and then identifying all other persons (eligible non-evicted) that were also eligible on those cases, and then generating the unique list of all eligible persons. Data primarily is for the period from 2004 to November 2016 as applicable.
ProgramEligibility	List of eligible persons and associated cases with public assistance programs including health care. This includes two de-identified encrypted unique ids: E_PMI (eligible person) and E_CaseNumber (eligible case the eligible person is associated with for an assistance program).
ProgramPayments	List of program payments [non-health care] made for those eligible cases. This data is at the case level.

Sanctions	List of all sanctions imposed on those eligible persons
FSSIndicator	List of all those identified as a "Family Stabilization Services" (FSS) status person for those eligible persons. FSS is a state-funded employment service track designed to serve participants who are at risk of long term welfare dependency due to employment barriers. FSS allows employment service providers more flexibility to develop appropriate plans based on the participant's individual circumstances.
PersonDemographicsOne	List of certain person characteristics [time variant] for those eligible persons: EducationLevel, MaritalStatus, USCitizen
PersonDemographicsTwo	List of certain person characteristics [time variant] for those eligible persons: Immigration status, Nationality
PersonDemographicsThree	List of certain person characteristics [time variant] for those eligible persons: Disability status
PersonIncomeJobs	List of income from jobs for those eligible persons
PersonIncomeUnearned	List of income from unearned sources: Child Support, SSI, and Other Unearned for those eligible persons
PersonIncomeBusiness	List of income from business (or self-employed) for those eligible persons
PersonExpenseRent	List of expenses for rent for those eligible persons
PersonExpenseUtility	List of expenses for utility: air, sewer, electric, fuel, water, other, and garbage for those eligible persons
PersonRelationship	List of relationships for those eligible persons associated with eligible cases
CaseAddress	Cannot be provided due to HIPPA data release restrictions.
SSIS	List of workgroups opened in Social Services Information System (SSIS) for those eligible persons
Shelter	The dataset includes shelter stays in Hennepin County shelters from January 2008 through December 27, 2017. Two sources were used- the first source Shelter Snapshot included family stays from 2008 to October 2, 2017 and single adult stays from 2008 to October 20, 2016. Single adult stays from this source were limited to 75% of total beds. The Homeless Management Information System (HMIS) was used for family stays from October 3, 2017 to December 27, 2017 and for single adult stays from October 21, 2016 to December 27, 2017. Single adult stays from this source were comprehensive. All stays have been collapsed into treatment and control categories comprehensive of single adults and families. Family stays were mostly entered into the data sources with the name of a head of household. It is possible for there to be children or multiple adults with their own lines if multiple people were entered for one stay.

FHPAP	List of persons who used the Family Homelessness Prevention and Assistance Programs (FHPAP) prevention funds anytime during July 2013 through January 2018. Included is the Person's first date of use of prevention funds during the time period.

7.2 Feature Importance

Base Data	Feature Definition	Coefficient Direction	Importance	Significance	Type
Current_Rent	Current max Rent Expense for a case in that month	increase	100	Important	Numerical
current_avg_age	Calculation of the mean age for all people currently on the Case	Increase	67.1304905	Important	Categorical
Months_on_FS	Cumulative months on Food Support for that case up to that date	Decrease	58.4735736	Important	Cumulative
non_english	Indicator if there is a non-english speaker present in the case at the current time	Decrease	56.8324854	Important	Categorical
Previous_denials	Cumulative previous denials for either emergency assistance or emergency general assistance over the entire data period	Increase	52.2656257	Important	Change
Months_on_HC	Cumulative months on Health Care for that case up to that date	Decrease	51.924112	Important	Cumulative
cur_months_on_HC	Indicator of if the case currently on a Health Care program	Increase	49.7663029	Important	Categorical
curent_case_benefits	Current case benefits for a case in that month	Decrease	49.4946953	Important	Numerical
ProgramPayment_Ratio	Current Program Payments / Max Program Payments for that case across all time	Decrease	47.9137042	Important	Ratio
RentRatio	Current Max Rent In Household / maximum rent for that case across all time	Increase	46.5050135	Important	Ratio
cur_months_on_FS	Indicator of if the case currently on a Food Support program	Increase	43.3777181	Important	Categorical

cur_months_on_MF	Indicator of if the case currently on a MFIP program	Increase	40.292171	Important	Categorical
females	Number of females currently in the household	Increase	36.8837261	Important	Numerical
males	Number of males currently in the household	Increase	34.8631101	Important	Numerical
IncomeRatio	Current Max Income in House Hold / Maximum income for that case across all time	Decrease	34.4684669	Important	Ratio
Current_ly_diff	Difference in household size when compared with the same case 6 months prior	Increase	28.9638937	Important	Change
Current_income_Jobs	Current max jobs income for a case in that month	Decrease	27.8466542	Important	Numerical
IncomeRatio_jobs	Current max jobs income in Household / maximum jobs income for that case across all time	Decrease	27.3464606	Important	Ratio
Current_income_other	Current max other income for a case in that month	Decrease	27.2084043	Important	Numerical
Change_FS	Change in Food Support for the case compared with 6 months prior. See change variables in section 3.3.5 for more details on calculation.	Decrease	25.2150669	Important	Change
Months_on_MF	Cumulative months on MFIP for that case up to that date	Increase	24.2239514	Important	Cumulative
HH_ratio_max	Current house hold size / maximum household size for that case across all time	Increase	23.8926822	Important	Ratio
HH_ratio_min	Current house hold size / minimum household size for that case across all time	Decrease	23.14178	Important	Ratio
Months_on_EA	Cumulative months on Emergency Assistance for that case up to that date	Increase	23.0111288	Important	Cumulative

DisabilityY	Does anyone on the current case have a disability (Y if so, N if not)	Increase	22.9140873	Important	Categorical
Current_income_SSI	Current max SSIS income for a case in that month	Decrease	21.4368823	Important	Numerical
Months_on_GA	Cumulative months General Assistance for that case up to that date	Decrease	21.3226653	Important	Cumulative
cur_months_on_GA	Indicator of if the case currently on a General Assistance program	Increase	21.2269924	Important	Categorical
IncomeRatio_unearned	Current Max Income (Unearned) in House Hold / Maximum income for that case across all time	Increase	20.4337759	Important	Ratio
Us_citizenY	Is everyone on the current case a US Citizen (Y if so, N if not)	Increase	18.4853725	Important	Categorical
Total_shelter_days	Cumulative days spent in a shelter for that case up until that month	Increase	17.2602391	Important	Cumulative
Change_HC	Change in Health Care for the case compared with 6 months prior. See change variables in section 3.3.5 for more details on calculation.	Increase	17.2388314	Important	Change
Months_on_EG	Cumulative months on Emergency General Assistance for that case up to that date	Increase	16.5210874	Important	Cumulative
Months_on_DW	Cumulative months Diversionary Work Program for that case up to that date	Decrease	16.2416799	Important	Cumulative
current_income_child_sup	Current max child support income for a case in that month	Decrease	15.0812852	Important	Numerical

Months_on_MS	Cumulative months on Minnesota Supplemental Assistance for that case up to that date	Increase	14.5233921	Important	Cumulative
Change_MF	Change in MFIP for the case compared with 6 months prior. See change variables in section 3.3.5 for more details on calculation.	Decrease	13.2795568	Important	Change
SSIS_child_protective_services	Cumulative months with an interaction with SSSIS program Child Protective Services for that case up until that month	Increase	12.67634	Important	Cumulative
Change_GA	Change in General Assistance for the case compared with 6 months prior. See change variables in section 3.3.5 for more details on calculation.	Increase	10.3613575	Important	Change
cur_months_on_MS	Indicator of if the case currently on a Minnesota Supplemental Assistance program	Increase	10.0183949	Important	Categorical
IncomeRatio_Bus	Current max business income in Household / maximum business income for that case across all time	Increase	9.59150909	Important	Ratio
Change_EA	Change in Emergency Assistance for the case compared with 6 months prior. See change variables in section 3.3.5 for more details on calculation.	Increase	9.58050974	Important	Change
Change_MS	Change in Minnesota Supplemental Assistance for the case compared with 6 months prior. See change variables in section 3.3.5 for more details on calculation.	Decrease	7.95854176	Important	Change

current_ssis_child_protective_services	Indicator of if anyone in the case currently on a Child Protective Services SSIS program	Increase	7.72614092	Important	Categorical
cur_months_on_EA	Indicator of if the case currently on an Emergency Assistance program	Increase	7.1091095	Important	Categorical
SSIS_child_welfare_gen	Cumulative months with an interaction with SSSIS program Child Welfare General for that case up until that month	Decrease	6.92912329	Important	Cumulative
SSIS_adult_services_gen	Cumulative months with an interaction with SSSIS program Adult Services General for that case up until that month	Increase	6.37346232	Important	Cumulative
Change_DW	Change in Diversionary Work Program for the case compared with 6 months prior. See change variables in section 3.3.5 for more details on calculation.	Decrease	6.17391664	Important	Change
Change_child_protective_services	Change in SSIS Child Protective Services program for the case compared with 6 months prior. See change variables in section 3.3.5 for more details on calculation.	Increase	6.16633935	Important	Change
Current_income_Bus	Current max business income for a case in that month	Increase	5.89621879	Important	Numerical
cur_months_on_GR	Indicator of if the case currently on a Group Residential Housing program	Decrease	5.72776217	Important	Categorical
cur_months_on_EG	Indicator of if the case currently on a Emergency Assistance General Work program	Increase	5.02915036	Important	Categorical

Change_EG	Change in Emergency General Assistance for the case compared with 6 months prior. See change variables in section 3.3.5 for more details on calculation.	Not Important	4.95792489	Not Important	Change
change_EG	Change in Emergency General Assistance for the case compared with 6 months prior. See change variables in section 3.3.5 for more details on calculation.	Not Important	4.95792489	Not Important	Change
Change_GR	Change in Group Residential Housing for the case compared with 6 months prior. See change variables in section 3.3.5 for more details on calculation.	Not Important	4.87564514	Not Important	Change
SSIS_child_mental_health	Cumulative months with an interaction with SSSIS program Child Mental Health for that case up until that month	Not Important	4.84660244	Not Important	Cumulative
months_on_WB	Cumulative months on Works Benefit Programs for that case up to that date	Not Important	4.58239068	Not Important	Cumulative
Current_shelter_days	Number of days spent in a shelter in the current month	Not Important	4.56706534	Not Important	Numerical
current_ssis_adult_services_gen	Indicator of if anyone in the case currently on a Adult Services General SSIS program	Not Important	4.52649712	Not Important	Categorical
SSIS_adult_mental_health	Cumulative months with an interaction with SSSIS program Adult Mental Health for that case up until that month	Not Important	4.13326114	Not Important	Cumulative

Change_adult_services_gen	Change in SSIS Adults Services General program for the case compared with 6 months prior. See change variables in section 3.3.5 for more details on calculation.	Not Important	3.33583674	Not Important	Change
cur_months_on_DW	Indicator of if the case currently on a Diversionary Work program	Not Important	3.3172347	Not Important	Categorical
current_ssis_adult_mental_health	Indicator of if anyone in the case currently on a Adult Mental Health SSIS program	Not Important	3.28860879	Not Important	Categorical
SSIS_chem_dependency	Cumulative months with an interaction with SSSIS program Chemical Dependency for that case up until that month	Not Important	3.2585257	Not Important	Cumulative
current_ssis_child_welfare_gen	Indicator of if anyone in the case currently on a Child Welfare (General) SSIS program	Not Important	3.18676382	Not Important	Categorical
SSIS_comm_alt_for_disable_indv	Cumulative months with an interaction with SSSIS program Community Access for Disability Inclusion for that case up until that month	Not Important	2.95650084	Not Important	Cumulative
change_child_welfare_gen	Change in SSIS Child Welfare (General) program for the case compared with 6 months prior. See change variables in section 3.3.5 for more details on calculation.	Not Important	2.51084686	Not Important	Change
Change_WB	Change in work benefits program for the case compared with 6 months prior. See change variables in section 3.3.5 for more details on calculation.	Not Important	2.38080884	Not Important	Change
cur_months_on_WB	Indicator of if the case currently on a Work Benefits program	Not Important	2.2048188	Not Important	Categorical

current_ssis_child_mental_health	Indicator of if anyone in the case currently on a Child Mental Health SSIS program	Not Important	2.06919642	Not Important	Categorical
SSIS_Other	Cumulative months with an interaction with SSIS programs with minmal variance in our dataset (**including** _____) for that case up until that month	Not Important	2.01499405	Not Important	Cumulative
SSIS_child_care_general	Cumulative months with an interaction with SSSIS program Child Care General for that case up until that month	Not Important	1.98707268	Not Important	Cumulative
current_ssis_comm_alt_for_disable_indv	Indicator of if anyone in the case currently on an Community Access for Disability Inclusion SSIS program	Not Important	1.77159923	Not Important	Categorical
current_ssis_child_care_general	Indicator of if anyone in the case currently on a Child Care General SSIS program	Not Important	1.70191347	Not Important	Categorical
Change_child_mental_health	Change in SSIS Child Mental Health program for the case compared with 6 months prior. See change variables in section 3.3.5 for more details on calculation.	Not Important	1.61059064	Not Important	Change
Change_chem_dependency	Change in SSIS Chemical Dependency program for the case compared with 6 months prior. See change variables in section 3.3.5 for more details on calculation.	Not Important	1.39239657	Not Important	Change
Change_adult_mental_health	Change in SSIS Adult Mental Health program for the case compared with 6 months prior. See change variables in section 3.3.5 for more details on calculation.	Not Important	1.31826684	Not Important	Change

SSIS_minor_parents	Cumulative months with an interaction with SSSIS program Minor Parents for that case up until that month	Not Important	1.09269523	Not Important	Cumulative
current_ssis_chem_dependency	Indicator of if anyone in the case currently on a Chemical Dependency SSIS program	Not Important	1.06953141	Not Important	Categorical
Change_child_care_general	Change in SSIS Child Care General program for the case compared with 6 months prior. See change variables in section 3.3.5 for more details on calculation.	Not Important	0.75182418	Not Important	Change
Change_comm_alt_for_disable_indv	Change in SSIS Community Access for Disability Inclusion program for the case compared with 6 months prior. See change variables in section 3.3.5 for more details on calculation.	Not Important	0.74007683	Not Important	Change
Change_other	Change in SSSIS programs with minimal variance in our dataset (**including** _____) for 6 months prior. See change variables in section 3.3.5 for more details on calculation.	Not Important	0.42291559	Not Important	Change
current_ssis_other	Indicator of if anyone in the case is currently on a other **(including)** SSIS program	Not Important	0.40419072	Not Important	Categorical
Months_on_4E	Cumulative months on IV-E Foster Care for that case up to that date	Not Important	0.27899278	Not Important	Cumulative
months_on_RC	Cumulative months on RC for that case up to that date	Not Important	0.26966864	Not Important	Cumulative

change_4E	Change in IV-E Foster Care for the case compared with 6 months prior. See change variables in section 3.3.5 for more details on calculation.	Not Important	0.07280923	Not Important	Change
cur_months_on_4E	Indicator of if the case currently on a IV-E Foster Care program	Not Important	0.04717541	Not Important	Numerical
current_ssis_minor_parents	Indicator of if anyone in the case is currently on a Minor Parents SSIS program	Not Important	0.01126427	Not Important	Categorical
Current_max_ed (ALL)	Ordinal sorting of education. The higher the education, the higher the number (ending at 18)	Not Important	0	Not Important	Categorical
cur_months_on_RC	Indicator of if the case currently on a Refugee Cash Assistance program	Not Important	0	Not Important	Categorical
Change_ssis_minor_parents	Change in SSIS Minor Parents program for the case compared with 6 months prior. See change variables in section 3.3.5 for more details on calculation.	Not Important	0	Not Important	Change

*Other SSIS includes Adult Essential Community Support, Child Brain Injury, Community Alternative Care Adult, Adult Foster Care, Child Alternative Disabled Individual, Alternative Care Waiver, Elderly Waver, Traumatic Brain injury, Child Care License and Early Intervention. These were bucketed together beause they had minimal variance (less than a thousand instances across the entire million plus record dataset).

7.3 Miscellaneous Analysis – Survival Models

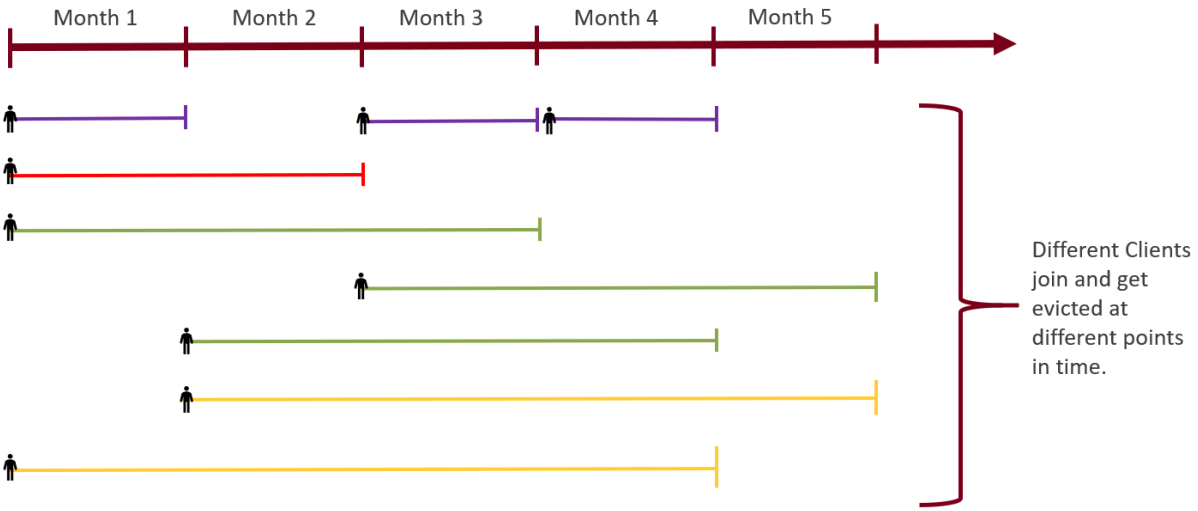
7.3.1 Introduction

Survival analysis models attempt to remove the temporal component of data in favor of accurate and descriptive predictions. In this analysis, we will look at how the "survival" rate of the population varies across different indicators. We define survival rate to be the proportion of the population that does not get evicted at a given point in time. For example, a 70% survival rate represents a population wherein 30% of the clients are evicted.

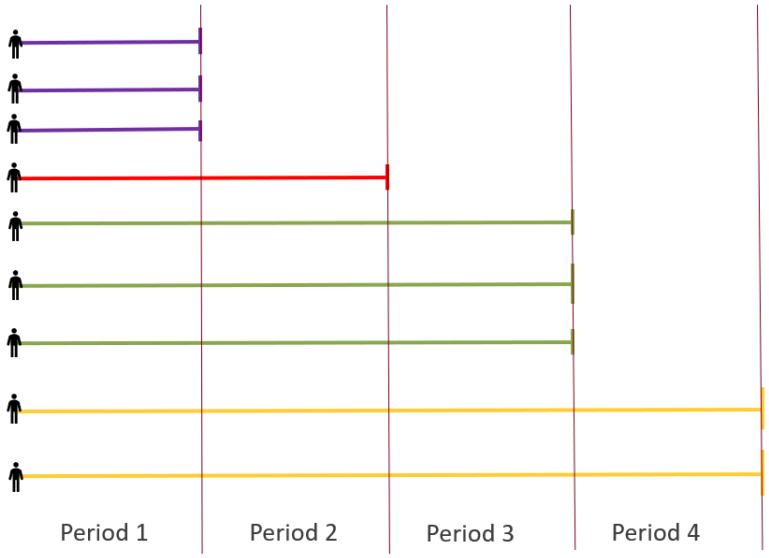
This definition helps us better understand which flags in the data are efficient at differentiating between potential evictees and non-evictees.

7.3.2 Model Intuition

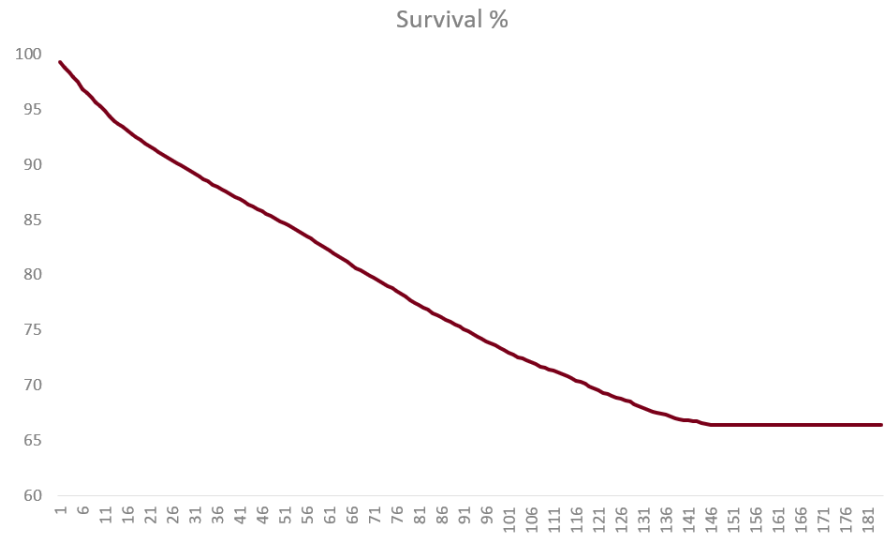
Assuming clients in the database appear at different points in the data and disappear after random intervals in time, we can plot a diagram like below:



We can remove the time element from the data by lining up all clients by start date and produce a diagram like below:



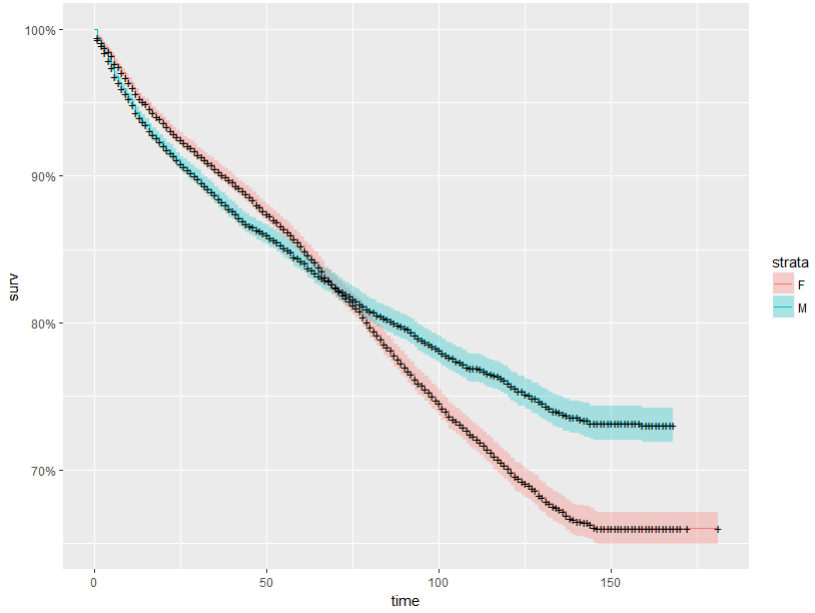
We can then plot an overall survival curve by counting an fraction of eviction with cumulative sum of periods.



This type of graph can similarly be plotted for different levels of categorical variables.

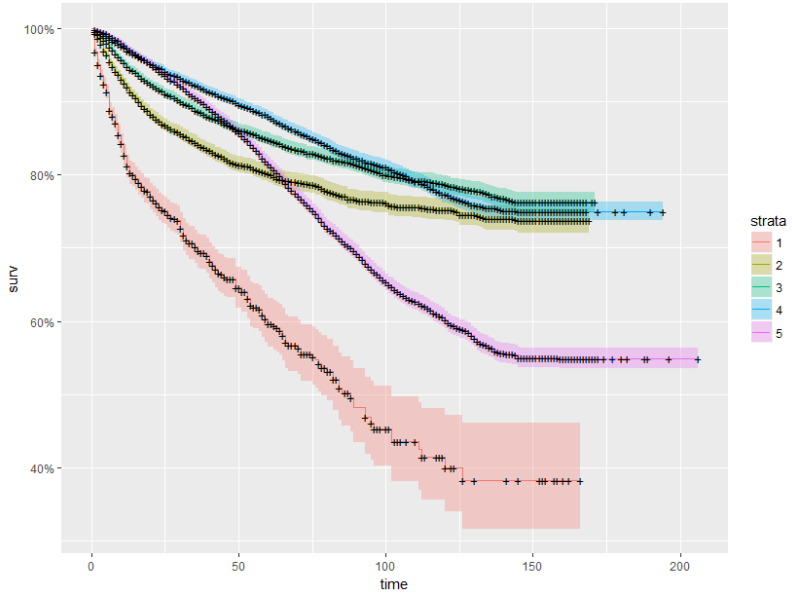
7.3.3 Preliminary Graphs

Comparison by Gender



The red curve signifies survival rate for women while the green one corresponds to men. We see that while during the first 70 months (periods), women are just as likely as men to be evicted, but after this point in time, the survival curve for females falls away steeply indicating women who stay for longer in the system are likely to get evicted quicker than men.

Comparison by Program Count



The curves in this graph represents the different number of programs a person can be a part of during his/her tenure in the dataset. Here we see that being part of just 1 program is just as

detrimental to a client as when the client is part of 5 or more programs. A sweet spot of 2 to 4 programs appears to be frequent for clients who are less likely to be evicted.

7.3.4 Model formulation

These univariate curves can be extrapolated further into a multivariate model that can attempt to identify characteristics that can differentiate evictees and non-evictees. Code that consolidates data for the multivariate survival model including test/train splits and the modelling functions are available at Survival Analysis.Rmd.

The confusion metric for test data used in the code is reproduced below.

		Ground Truth	
		0	**1**
Reality	**0**	3239	2333
	1	530	**705**

As we can see here, we achieve a false positive rate of 57% with an accuracy of 58%. Though these are large numbers, the time-sensitive nature of the data has been discarded in favor of identifying features that are effective in flagging evictees. While our attempt at making this model time-sensitive was unsuccessful, the knowledge we gained from this model informed us in our development of the more conventional models discussed in earlier sections.