Methods of Imputation and Data Merging to Predict Supportive Housing Outcomes for Homeless Families in San Francisco

Methods of Imputation and Data Merging to Predict Supportive Housing Outcomes for Homeless Families in San Francisco

A THESIS PRESENTED
BY
MADELEINE R. NAKADA
TO
THE DEPARTMENT OF COMPUTER SCIENCE AND THE DEPARTMENT OF STATISTICS

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
BACHELOR OF ARTS
IN THE SUBJECT OF
COMPUTER SCIENCE AND STATISTICS

HARVARD UNIVERSITY
CAMBRIDGE, MASSACHUSETTS

APRIL 2020

Thesis advisor: Dr. Pavlos Protopapas Madeleine R. Nakada

Methods of Imputation and Data Merging to Predict Supportive Housing Outcomes for Homeless Families in San Francisco

Abstract

Machine learning models have been applied to data on homeless households to investigate a range of problems from predicting length-of-stay at homeless shelters to predicting outcomes for homeless individuals to help cities and shelters allocate resources. However, the majority of these models rely on data from collected from a number of agencies to build a comprehensive dataset of homeless individuals within a municipality. Fitting a model to predict outcomes for households on data from a single provider presents a number of challenges including a smaller dataset and a lack of secondary data sources to deal with missing data. Nonetheless, such a model can be useful to these service providers as it can be used to model how households respond to the provider's specific services rather than generalizing over all supportive housing providers within a city who may cater to different demographics and have different levels of engagement with In this thesis, I investigate methods for building a dataset which can be used to predict outcomes for households that engage with Compass Family Services, an agency which provides housing and housing stipends to homeless households in San Francisco. The results show that while relatively high prediction accuracy can attained using simple imputation methods, these accuracies rely on information in the dataset that includes information about their enrollment in other programs. Since these programs filter who can enroll in them, these variables are likely correlated with other agency's beliefs that the household will have a successful exit. I discuss the pros and cons of including these variables, as well as further applications of the predictive model to chronic homelessness.

Contents

Listing of figures

This is dedicated to my parents.

Acknowledgments

First, I'd like to thank Kendra Froshman and Compass Family Services for allowing me to use their data and to learn about their organization. I'd also like to thank them for all the that they do to help support families in the San Francisco community.

Many thanks to my thesis advisor Pavlos Protopapas for pushing me to apply the skills from his classroom to my community. Thank you to Weiwei Pan for giving me direction when thing went awry and always being available to talk through a problem.

Thank you also to Nicole Kim keeping me sane with on our weekly runs along the Charles, and to Anli Chen and Clea Schumer for your unending optimism. And finally, thank you to Daniel Inge for endless support and snacks.

0

Introduction

The 2019 San Francisco Homeless Point-in-Time Count & Survey found that during a single night in January 2019, 8,011 individuals were experiencing homelessness in the city. Of these homeless individuals, roughly 3,028 or nearly 38% were classified as chronically homeless, having been continuously homeless for one year or more or experiencing at least four separate episodes of homelessness totaling 12 months over the past 3 years ([6]).

Much of the existing research around applications of machine learning to the issue of homeless-

ness focuses on predicting outcomes for individual households based on interventions taken to help them obtain temporary or permanent housing ([4], [2]). While these models can be useful in resource allocation or when trying to maximize the number of positive exits from a program, identifying families that are less likely to succeed in a program or are more likely to return might allow an agency to allocate more resources towards these households to increase their chances of success.

While it may seem counter intuitive to allocate more resources to households that are less likely to obtain permanent housing, the short term cost of these additional resources may outweigh the long-term costs of supporting a chronically homeless household. In a study of two subsets of the chronically homeless population–chronically homeless individuals who had lost their jobs and chronically homeless youth–Toros et al. found that health care costs for the two populations were respectively five and four times higher than a comparable non-homeless demographic [8]. Justice system costs were nine times and seven times higher for the two populations as well. Furthermore, while costs for non-homeless and short-term homeless households declined over the course of the three-year study, costs for chronically homeless households did not, widening the gap in costs between the chronically homeless and the rest of the population.

Hong et al. address chronic homelessness in their 2018 study of households in shelters in New York City run by the non-profit Women in Need [3]. The paper focuses on predicting which households will re-enter the shelter system after initially leaving. The authors use demographic data combined with information about the cause of a household's homelessness to estimate the probability that a household will re-enter the system. Although the authors found that their model had fairly strong predictive power, the conclusions of the paper focused on which parameters of the model were the most significant in determining the probability of re-entry.

Motivated by the Hong et al. [3] study, I have carried out a similar analysis on homelessness in San Francisco. However, whereas Hong et al. incorporate data from multiple city agencies to supplement the WiN data, this study uses data from a single agency in San Francisco. Due to the reduced

size of this dataset, the first half of the study focuses on methods for building a robust dataset to be used for prediction. I then apply a similar approach to the WiN study both to predict outcomes for homeless households enrolled in supportive housing programs and to predict which households are likely to re-enter the system after exiting. The goal of this study, as with the Hong et. al study, is not to create a model that can be put into practice determining who gets access to services and who doesn't, but rather to identify what features make a household more or less likely to succeed.

1

Background

Compass Family Services is a non-profit organization in San Francisco that provides both short and long-term housing and housing support to homeless families. Compass also provides these families with childcare, childhood education, referrals to mental health support, and job readiness preparation.

In 2017, Compass served 5,143 individuals, 52% of which were children under the age of 18. Compass reports that 94% of families that complete their housing programs maintain stable housing

after program completion. It should be noted that although Compass has an emergency shelter, the programs which I've focused on are not emergency housing services, and Compass screens households before accepting them into the program.

In my research, I focus on three of Compass's main programs to help families that are housing insecure.

- Compass Connecting Point: CCP serves as an entry point for Compass's other supportive housing programs. Clients that go through CCP can be referred to other Compass programs or shelters and housing programs outside of Compass
- Compass SF Home: CSFH is a rapid rehousing program that gives families rental subsidies. Families in CSFH receive case management one per month
- Compass Clara House: CCH is a two year residential program which provides housing to families. Families receive weekly case management. The program is limited to 13 families at a time.

Compass provided four anonymized datasets with data collected from October 2011 through January 2020 for clients in these three programs. Households were given a unique identifier linking entries across datasets.

1.1 The Datasets

With the exception of the client dataset which includes demographic data collected at intake, the datasets include data collected over the course of a client's enrollment in each program. These measurements are taken in the form of assessments that are administered at case management programs, the frequency of which varies by program. A full list of the features of each dataset can be found in Appendix A.1.

1.1.1 Client Data

In the client dataset, one row represents one household's enrollment in a single program. Households may have multiple rows due to being enrolled in multiple programs, or the same program multiple times. The dataset includes the program name, open date and exit date of the client's case within the program, as well as demographic data about the household. Households are represented by a single head of household whose age, race, and gender is recorded in addition to the size of the household, and other programs that the household is enrolled in. There are 1,080 rows in the client dataset.

1.1.2 Income Data

The income dataset includes reported household income at case management sessions split between cash income and non-cash benefits. Cash income includes both income from jobs as well as cash benefits such as social security. Non-cash benefits include benefits such as food stamps. The dataset also includes the program in which the household was enrolled when the assessment was taken. There are 1,005 rows in the dataset.

1.1.3 Assessment Data

The assessment dataset includes assessments performed each quarter during case management sessions. The assessment rates households on a scale from 1-5 in areas such as mental health, interactions with Child Protective Services, food access, and employment. A score of 1 indicates a family in crisis whereas a score of 5 indicates a family is doing well. The dataset also includes a client's score on the Adverse Childhood Experience (ACE) questionnaire. This questionnaire is administered once, typically close to intake, and is intended to be a static score measuring an individual's childhood trauma. Scores range from 0-10 with 10 indicating the highest level of childhood trauma.

Assessment date and the program in which the household was enrolled at the time of the assessment are also included in the dataset. There are 1,434 rows in the dataset.

1.1.4 Housing Tracker Data

The housing tracker dataset includes information about households' living situation, enrolled program, type of subsidy or supportive housing. The dataset also includes the start date of the subsidy/supportive housing, as well as whether the assessment was taken at program intake, program exit, or during program enrollment. There are 2,524 rows in the dataset

2

Methods

This chapter is divided into three parts. First, I combine the individual datasets into a single dataset where each row represents a household's enrollment in one of Compass's program. Next, I use the dataset to predict outcome types for households enrolled in Compass Clara House and Compass SF Home. Using the insights from the outcome prediction model, I then predict the likelihood that a family that enrolls in a Compass program will later re-enter the Compass system.

2.1 Combining Datasets

The three datasets of interest, income data, client demographic data, and assessment data need to be joined such that each row in the new dataset represents a measurement at a single point in time for the same family. Combining the datasets would appear to be a straightforward process since all datasets contain a timestamp and share a unique identifier for each household. However, a significant amount of missingness arises when the datasets are merged using this timestamp. In addition to human error, many of the entries in the three datasets are not well-aligned due to differences around when an individual was technically enrolled in a program, and when they underwent their initial assessment to enroll in the program. In this section, I discuss methods for combining the datasets and how decisions surrounding the merging of these datasets may have introduced sources of error or noise in the data.

Merging Datasets

The client dataset serves as the basis for the complete dataset, meaning that the "Open Date" in the client dataset will be the ground truth on which dates in the other datasets are compared to. Only approximately 27% of the 1080 rows in the client dataset have an income measurement for a given household on the same day as a case was opened for that household. However, loosening the constraints to include any income measurement within 21 days of the client dataset Open Date increases this matching percentage to 40%. A further decrease in threshold to match any income measurement within 90 days increases the percentage of client data rows with matching income to almost 49% or 526 entries.

However, there's a trade-off between decreasing the threshold to get a larger matched dataset, and an increased likelihood that the individual's income has changed between the time they started the program and the time their income was measured . Fig. 2.1 shows how the distribution of the

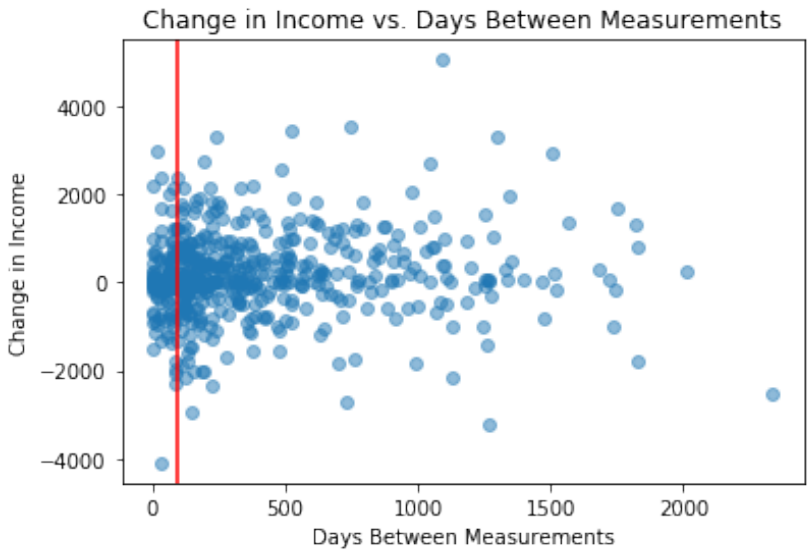

Figure 2.1: Plotting Days Between Income Measurements vs. Change in Income

difference between income measurements changes as the time between measurements increases. For the purpose of analyzing changes in income, cash and non-cash benefits income were combined to calculate total income. This distribution was found by taking families that had multiple income measurements within a period of time less than the threshold and calculating the difference.

I use a threshold of 60 days for an income measurement to qualify as "matching" a household's enrollment in a program. On average, households with multiple income measurements within a 60 day period saw a change in their total income although nearly half of households saw this change as an increase and half saw their income decrease. This potential discrepancy between measured incomes which are "matched" within the 60 day window and a household's true income on the day of enrollment is a potential source of noise which could skew the accuracy of predictions mae on this data

The combined income and client data dataset is created by adding the corresponding cash income and non-cash benefits from the matched income rows to the client rows. The combined in-

come and client dataset must then be merged with the assessment dataset. There are similar date misalignments between the assessment and the combined dataset. I use the same method of increasing the range of dates that are considered to be a match to match them.

Only 79 of the 483 entries in the combined dataset, or 16.4%, have an assessment within 21 days of the program's start date. Lowering the threshold to 90 days increases the number of matches to 147. Completely ignoring the assessment date, 374 entries (77.4%) in the combined dataset have a matching assessment at any point in time, but the average number of days between enrollment and assessment date is 295 with a standard deviation of 390 so it's very likely that the assessment will be out of date.

There's almost a doubling in the number of matched entries when the threshold is increased from 21 to 90 days. However, given the fact that some programs only enroll households for a couple of weeks, a household could go through a number of programs over the course of 90 days, resulting in a very different assessment score than when they enrolled in the first program. To account for the differences in program length, the threshold is set to match any assessment score taken during the time that a household was in the given program. This means for a family enrolled in a program for a week, a matching assessment must fall within the week they were enrolled in the program. A family enrolled in a program for a year can have a matching assessment that falls anywhere within that year. If there are multiple matching assessments, the closest one to their start date is used. This leads to a match of 263 of the combined entries that have client data, income data, and assessment data all aligned by program start date. This small dataset will be called the Full Dataset.

The remaining entries in the client dataset that are missing matching assessment or income data will not be discarded. A separate dataset is created based off of the 1037 client entries. The aligned assessment and income data found above is added to this dataset, leaving rows with unmatched assessment data as missing values. This dataset will be called the Missing Dataset

Pre-Processing

One notable outcome of the initial exploratory data analysis of the dataset is that a handful of categorical demographic predictors occur very rarely. Having a predictor that only only occurs for one to two entries in the dataset runs the risk of having perfect separation in a training set created from that dataset since it's possible that all entries that have a particular categorical predictor also have the same response. Households with these demographic predictors were assigned either to similar predictors in the dataset, or categorized as "Other" for the category which the predictor was a factor of.

As a final step, the rows missing demographic data for race and gender are removed since these demographic variables may be harder to impute. After taking into account all merging, the Full Dataset has 46 predictors and 220 rows. Missing Dataset has 46 predictors and 1042 rows.

Creating a Response Variable

The variable of interest is the type of exit that a household has from a program. The "Exit Reason" column in the client dataset is a string selected from a finite subset of options that indicates why a client exited a program. There are 21 possible "Exit Reasons" and so the response variable could be a categorical response of these possible exits. However, in the Full Dataset, many of the Exit Reasons have only one or two households that exited with this reason and so, again, there is a risk of having perfect separation in the training set. Instead, I reduce the problem to a binary classification, classifying the 21 Exit Reasons as positive, negative or neutral as shown in Appendix A.2. Only the positive and negative exit reasons will be used in the model.

Neutral exits also include data without an exit reason. Almost half of entries in the client dataset are missing an exit reason which is a significant portion of the data. An exit reason may be missing due to errors in data input, an individual losing contact with Compass and not formally completing

their program, or an individual who is still active in the program and has not yet completed or exited from a program. Ignoring the neutral exits, the problem of predicting a household's outcome can be seen as a binary classification problem to predict success or failure within a program.

2.2 Predicting Outcome Type for Supportive Housing Programs

Preliminary Models

Before fitting and tuning more complex models to predict whether a household will have a successful exit from a given program, simple models can be fit to determine whether or not the given predictors in the dataset are significant predictors of the response variable.

The Full Dataset is suitable for this task since the dataset doesn't have data missing not at random as a result of unmatched income and assessment data like the Missing Dataset does. However, the Full Dataset does still have some data missing at random which can be attributed errors in data collection or data input where an assessment was completed, but data is missing for a subset of the assessment questions.

Since the initial model is just used to indicate if there is a signal from the predictors, a simple mean imputation can be used to impute the mean of each column onto its missing values. The mean imputation is implemented in a stratified manner by "Program Name" where, for each program, the mean for a particular predictor is calculated for the entries not missing a value for that predictor, and then imputed on the entries for households that are missing the predictor and were enrolled in the given program.

I use six common models for classification using commonly used parameters without tuning. The models are implemented using the scikit-learn library. The models are as follows:

- Logistic Regression: Logistic regression takes on the form

$$P(Y = 1) = \frac{1 \exp(\beta X)}{1 + \exp(\beta X)}$$

with regression coefficients β. The returned probability is the probability that the response variable is class 1, in this case, a positive response. Logistic regression can be used as a classifier by classifying a set of predictors as belonging to Class 1 if $P(Y = 1) > t$ for some threshold t. By default, $t = 0.5$ although it can be adjusted through hyperparameter tuning. Logistic regression was used by [3] in their predictions of re-entry to the shelter system. I hypothesize that re-entry to the shelter system and negative exits may be related, which is a motivation for including this model.

- Linear Discriminant Analysis: Given predictors x, let

$$f_k(x) = P(X = x | Y = k)$$

for $k = 0, 1$. That is, let $f_k(x)$ be the density function for X given that Y=k. Assume that the predictors are each independently normally distributed with equal variance. Using Bayes' Theorem, we can derive

$$P(Y = k | X = x) = \frac{\pi_k f_k(x)}{\sum_{t=1}^{K} \pi_t f_t(x)}$$

where π_k is the prior probability of class k. The class with the largest posterior probability for data x is used as the prediction

- Quadratic Discriminant Analysis: QDA follows the same process as LDA, but drops the assumption that the variances of each variable's distribution are equal.

- Gaussian Naive Bayes

- Multinomial Naive Bayes

Due to the small size of the dataset, simply measuring test accuracy is not necessarily an accurate measure of model fit since one or two outliers in the test dataset could significantly draw down the accuracy. To address this, test accuracy was measured using five-fold cross-validation.

The imbalance in the response variable resulted in many of the models misclassifying negative outcome households as positive outcome households. To remedy this imbalance, class weights are assigned to each class so that the negative class is more often predicted. Since this is just an initial model, even class weights are assigned to each class rather than tuning the class weights. The above models are re-fit with these weights.

2.3 Impute Missing Data

It's necessary to use the Missing Dataset to further investigate the prediction of Exit Scores since there is not sufficient data in the Complete Dataset to model Exit Scores for Compass Clara House. However, in order to use the Missing Dataset the missing values have to be imputed.

For the simplest form of imputation, I again use mean imputation, using separate means for each program for each predictor. This simple imputation serves as the baseline for more complex approaches.

Regression Imputation

In regression imputation, the missing data was modeled as a function of the other predictors. The Full Dataset has three levels of missingness. The first level, the demographic data, has no missing values since all missing rows have been removed. The second level, Total Cash Income and Total Non-Cash Benefits, are both missing in approximately 55% of rows. The third level, the Assessment Dataset measurements, are missing in 73-83% of rows.

First, I model Total Cash Income and Total Non-Cash Benefits as a function of the demographic data. This problem is a regression and so I use four common regressors:

- Linear Regression: $y = \beta X$
- Ridge Regression: Linear regression with a penalty on the square of the coefficients
- Lasso Regression: Linear regression with a penalty of the absolute value of the coefficients
- AdaBoost: AdaBoost is a variation on decision trees. With each iteration, the datapoints are re-weighted such that the misclassified datapoints are more heavilty weighted

For all four regressors, I use the default sklearn module with the intention of tuning the best model. However, based on the results of these preliminary models, there doesn't seem to be enough signal from the demographic data to predict income (discussed further in Section 3.2) and instead I impute the income predictors.

The same set of regression models is run for each assignment score column, with demographic and mean imputed columns as predictors. For each program, the model with the highest test score is selected. Models with a test score less than 0.1 are discarded, and mean imputation is used instead on the column. Models with a test score greater than 0.1 are used to predict and impute the remaining missing values.

MICE

I use multivariate imputation by chained equations (MICE) as a third method of imputation. This was implemented using the statsmodel library[7].

At a high level, MICE initially uses a simple imputation to fill missing values. Then for one column at a time, the missing values are removed. The posterior distribution of the remaining values

in the column as a function of the other parameters is used to impute the missing values in that column [5]. The statsmodel package assumes a gaussian posterior distribution.

The process is repeated for each column to obtain a single imputation. Multiple imputations can be obtained by repeating the process for each column, starting with the already imputed values from the previous iteration.

The statsmodel MICEData function throws an error when more than 70% of values are missing and so I drop all columns that are missing more than this threshold for a given program. I took the average over five MICE imputed datasets to obtain the final dataset on which the predictive models were fit.

2.4 Hyperparemeter Tuning on Larger Dataset

I replicate the procedure from the initial dataset on the imputed datasets, fitting the same five models. However, as in the initial dataset, these models have much higher training accuracy than test accuracy, indicating overfitting on the training set.

To address overfitting in the logistic regression model, I tune the regularization parameter for both L1 and L2 regularization. Whereas L2 regularization, the regularization used on the Full Dataset, penalizes the square of each coefficient in, L1 regularization penalizes the absolute value of the coefficient. For the LDA models, I test combinations of shrinkage and solver for each dataset as well. After fitting the imputed datasets on the above models, I add interaction terms between all variables and re-fit the models.

2.5 Predicting Return Rates

Having shown that the predictors in the dataset are sufficient to fit a model to predict whether a household will exit with a positive or negative exit score, I can now use the same methodology to

predict whether a household will return to Compass after exiting from a supportive housing program.

To predict return rates, I use mean imputation on the Missing Dataset since it had the best performance in previous models. A returned household is considered to be a household where the household was previously enrolled in either Compass Clara House or Compass SF Home and later had an intake taken at Compass Connecting Point. Out of households 157 household entries in the Compass SF Home, 30 of the households, or roughly 23.6%, returned to CCP after exiting. Based on the data, there is no way of knowing whether a family was homeless when they returned to CCP, or if they returned to get advice on how to maintain their housing or to receive additional services. In their 2017 annual report, Compass reported that 94% of households that complete their programs remain housed after. However, this dataset also includes households with negative exits which are not included in Compass's statistic[Com].

3

Results

3.1 Basic Models

The baseline model for predicting a household's Exit Score is to predict the majority class 100% of the time. Since the outcomes are so skewed, this actually sets a fairly high bar for a model to beat. 95.5% of households enrolled in Compass Clara House in the Full Dataset have a positive exit score. The data is slightly less skewed for Compass SF Home with 64.5% of households having a positive

exit score.

The Full Dataset contains 19 household enrollments in Compass Clara House and 45 household enrollments in Compass SF Home. The skew in the Compass Clara House dataset, combined with the small number of households enrolled in the program means that there is just one entry in the Compass Clara House subset of the Full Dataset with a negative exit, making it impossible to train and test models on the dataset. Instead, I focus on the Compass SF Home data in this section, and analyze Compass Clara House further on the Missing Dataset after running imputations.

For the Compass Clara House data, the logistic regression model achieves .86 accuracy on the test set, indicating that the parameters are strong predictors of the response variable. This is higher than the baseline accuracy of .645 from predicting the majority class.

Given the skew of the response variable, I also evaluate the misclassification rate for each class to gain insight into how the model is performing. One measurement for multi-class classification is the Area Under the Receiver Operating Characteristic Curve (ROC AUC). The ROC curve plots a model's true positive rate against its false positive rate. The ROC AUC is the area under this curve.

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{TN + FP}$$

A well-fit model should have an AUC close to 1, indicating that it can simultaneously attain both a high true positive-rate and a low false-positive rate. An AUC close to 0.5 indicates that a model is unable to discriminate between classes.

The logistic regression model on the CSFH dataset was found to have an average ROC AUC of 0.827 using five-fold cross validation. The ROC AUC is fairly close to 1, indicating that the model is able to discriminate between households with positive and negative exits.

The ROC AUC itself can be misleading due to the imbalance of classes in the dataset. Since the

	Normal Train	Normal Test	Balanced Train	Balanced Test	Skewed Train	Skewed Test
Logistic Regression	0.973333	0.861538	0.985	0.784615	0.982857	0.769231
LDA	1.000000	0.538462	1.000	0.569231	0.994286	0.569231
QDA	1.000000	0.476923	0.500	0.692308	1.000000	0.615385
GNB	0.873333	0.615385	0.885	0.615385	0.891429	0.584615
MNB	0.513333	0.384615	0.590	0.384615	0.688571	0.384615
Adaboost	0.840	0.723077				

Figure 3.1: Test Set accuracies with varying degrees of upsampling of the minority class

response variable is skewed towards successful exit, the AUC may be high despite a fales negative rate becuse a small number of true negative cases means that a model can minimize the false-positive rate by under-predicting the negative class. In this case, looking at the raw confusion matrices can also be helpful.

Looking at the summed confusion matrices from the five-fold cross-validation the logistic regression model, the accuracy rate for households with positive exits, or the recall, is .867. However, for households with negative exit rates, the accuracy rate, or specificity falls to 0.65.

The low accuracy score for the minority class isn't surprising given the imbalance in the data. To lessen the effect of this imbalance, I re-fit the logistic regression model on two datasets with the minority class upsampled. In one dataset, the balanced dataset, the households with negative exit scores are upsampled so that there are an even number of positive and negative exits in the dataset. In the second dataset, the skewed dataset, I weight the classes so that they are inversely proportional to their true weights. That is, given n_p households with positive exits and n_n households with negative exits, I upsample such that

$$n_u = \frac{n_p}{n_n} n_p$$

where n_u is the number of households in the upsampled class (rounded).

3.1

As the degree of upsampling increases from none to biased towards the minority class, the specificity increases from 0.65 to 0.75, which is the desired outcome. However, with the increase in specificity there is also a decrease in recall from 0.867 to 0.75. Since the data is upsampled only in the training data, the effect of the increase in specificity is much smaller on the total accuracy score is much smaller than the effect of the decrease in recall and the overall test accuracy decreases as the proportion of negative outcome households in the training set increases.

In practice, if this model were to be implemented, the recall and specificity would be tuned to fit the respective costs of false positives and false negatives. However, since the goal here is to find evidence that the predictors can be used to predict the outcome variable, I will use un-upsampled data moving forward so that the accuracies can be compared between models.

3.2 Regression Imputation

Predicting Income

In this section I include results for Compass Connecting Point since I use imputed data for CCP households later in my analysis. The sklearn library uses the coefficient of determination of the squared residual (R^2) as its scoring function for most regressors. A model with an R^2 of 1 on the test data is able to explain 100% of the variation in the data. A model with an R^2 of 0 on the test data performs as well as simply predicting the mean of the data. A model with a negative R^2 performs worse than predicting the mean.

All of the models have negative R^2 scores when predicting Total Cash Income on the test data, indicating that the set of demographic predictors is not sufficient to explain variations in Total Cash Income. In fact, predicting the mean would perform better than these models.

For Total Non-Cash Benefits, CCP still has negative scores for all the models on the test set.

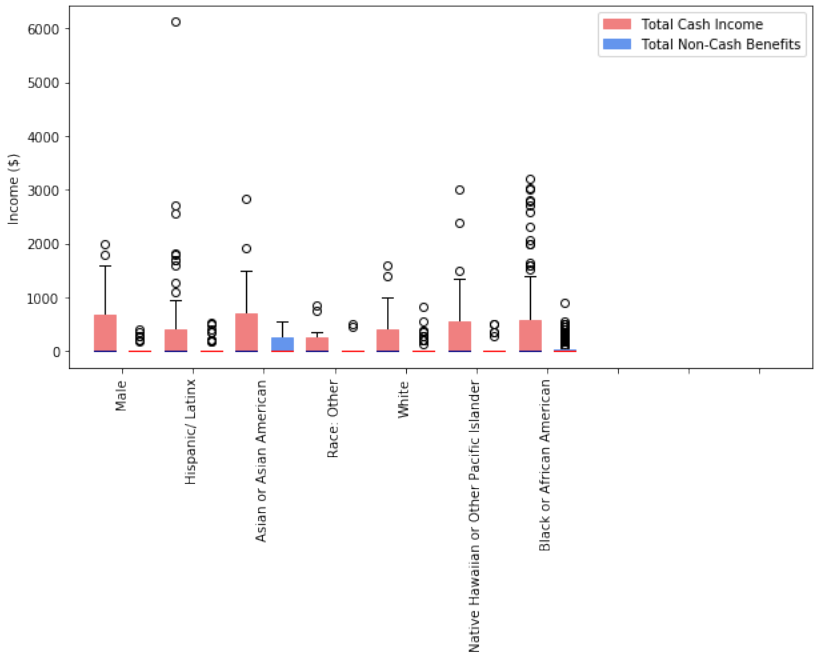

Figure 3.2: Distribution of Cash and Non-Cash Income by Demographic Predictor

CCH has one positive score of 0.084274 on the ridge regression model, but the score is still close to 0, indicating that the model is not much better than just predicting the mean. The same is true for CSFH which has all of its test scores greater than 0, but still around 0.1.

The lows scores are not that surprising given the initial exploratory data analysis on the relationship between income and the demographic predictors in the dataset. Fig. 3.2 shows that by visual inspection there is not a significant difference in the distribution of incomes for households between demographics.

Predicting Assessment Scores

For each assessment column, I fit the four models on the demographic and income data with the column of interest as the response variable. For the Compass SF Home data, every assessment column has at least one model with an R^2 score greater than 0.1. Compass Clara House has 13 out of

15 columns with a model with an R^2 score greater than 0.1, and Compass Connecting Point has 6. I mean impute the data for columns without a model above the 0.1 threshold.

3.3 Comparison of Predictive Power of Imputed Datasets

None of the models perform better than the baseline for Compass Clara House. A number of the models across all datasets have an accuracy score of 0.916667. However, looking at the confusion matrix for these models over a number of cross-validations, it becomes evident that these models are predicting positive outcomes for every household in the test sets.

Upsampling the households with negative outcomes as I did with the initial dataset fails to improve the specificity of the model. In some cases, for example with the logistic regression model on the mean imputed data, upsampling results in a lower recall rate, but still 0 specificity.

On the Compass SF Home Data, all three imputed datasets fit models that perform better than the baseline. Overall, the best performing model is logistic regression tuned to have a regularization coefficient of 0.2 with L2 error and fit on the mean imputed data with interaction terms. With an accuracy score of 0.706383, this model performs slightly better than the best model from the dataset imputed with regularization.

Comparing the misclassification rates of the top models from each dataset, the LDA model fit on the regression imputed dataset has slightly higher AUC than the logistic regression model fit on the mean imputed dataset. Both have an AUC above .77, indicating that they're strong predictors of the response variable. The AUC for the model fit on the MICE imputed data is closer to 0.5 at 0.57.

Evaluating the true positive and true negative rates, the model fit on the regression imputed data has both higher true positive and true negative rates, compared to the equivalent rate for the mean imputed dataset. Furthermore, the two rates are nearly equal meaning the model has similar accuracy on households with positive outcomes as it does on households with negative outcomes.

Although the LDA model on the regressed dataset has higher AUC, I use the logistic regression model fit on the mean imputed data with interactions to analyze the effects of various predictors on a household's probability of having a positive exit score.

The largest coefficient for any of the predictors has a magnitude of 0.06, which is very close to 0. A potential cause for the small magnitude of the data is the fact that there are significantly more predictors than rows in the dataset as a result of taking all the interaction terms. This means that there isn't a unique solution to optimizing the coefficients of the predictors, which explains the unexpectedly small coefficients.

The second best performing logistic regression model does not have interaction terms and has coefficients of reasonable magnitude. However, the top two coefficients are for enrollment in RAP and AFT. RAP is the State Rental Allowance Program, a type of housing subsidy. The fact that these two predictors are significant doesn't necessarily indicate that enrollment in these programs will lead to an increased likelihood of positive outcomes since it's likely that households that are enrolled in these programs are selected based off some criteria.

Evaluating the quantitative value of coefficients in logistic regression is difficult because an increase in the log-likelihood of an outcome is not an easily interpretable measurement. However, I can evaluate the comparative magnitudes of the coefficients as well as their signs. Since I use regularization in this model, the majority of the coefficients have been set to 0, leaving a subset of significant coefficients.

The largest magnitude predictor following the program predictors is family size which appears to be negatively related to the response variable, indicating that larger families are less likely to remain out of Compass's system after exiting. The two remaining predictors, Asian or Asian-American and Age both have positive coefficients although their magnitude is quite small, meaning that their contribution to the likelihood of a positive exit might be negated by the Family Size coefficient.

Removing these potentially confounding predictors and re-fitting the model results in a de-

creased test accuracy of .65. However, this is still higher than the baseline so the coefficients are worth investigating. Of the significant predictors, only had two positive coefficients: Age (0.019) and Asian-American (0.31). The rest, ACCESS, Family Size, and Hispanic/Latinx all had negative coefficients.

3.4 Predicting Return Rates

The best model for predicting return rates is Linear Discriminant Analysis with shrinkage. It attains .81 accuracy on the test set. The logistic regression model performs nearly as well with .795 accuracy with the regularization constant set to 0.1. The most significant predictor in terms of magnitude is enrollment CCS, which is positively correlated with an individual re-entering the system. However, the next five of the top predictors are all demographic or assessment parameters. The full list of coefficients is listed in A.4.

4

Conclusion

4.1 Summary and Discussion

By far the most difficult part of this process was trying to build a dataset that I could make inferences on. Even obtaining the dataset took months of coordination and discussions, highlighting the need for methods to make predictions and inferences on a smaller subset of data than previous research has used since as the Department of Housing and Urban Development is increasingly cen-

tralizing data surrounding homelessness, this data is also getting more difficult for researchers to access.

In the end, the degree of missingness in the Missing Dataset that I built from the four Compass Datasets was too significant to model the missingness as a function of the other parameters and the mean imputed data overall performed best. There was further evidence that the degree of missingness was too substantial in the fact the the MICE imputation dataset performed significantly worse than mean imputation for the purpose of prediction.

Nonetheless, the best performing model on the CSFH data had 0.706383 accuracy, showing that even a dataset with a significant degree of missingness can be used for prediction and inference. It wasn't until I analyzed the coefficients of that best model that I realized that enrollment in these programs could be highly correlated with the response variable. However, even after removing these variables from the dataset, the model outperformed the baseline model with an accuracy of approximately 0.65.

Similar to Hong et al., I found logistic regression to be one of the best performing models in almost all circumstances. Although I began this project hoping to model homelessness as a reinforcement learning model, the constraints of the data forced me to re-evaluate my approach. However, when it comes to social problems like homelessness, where implementations of these models decides who gets resources and who doesn't the explainability of these simpler models in addition to their accuracy, is paramount.

4.2 Future Work

The next step in the process would be to do statistical inference on final logistic regression models to determine which parameters are significant beyond just an analysis of their magnitudes.

Another potential area of research is in the interaction terms. The initial model with interactions

terms performed well, but due to the size of the dataset it was unclear if this was due to the fit of the data or the model overfitting. I'd like to run feature selection on the polynomial models to see if a more robust model might be built with the interactions, but with fewer features.

Lastly, I didn't get a chance to use the Compass Connecting Point data but it's a unique dataset compared to most datasets relating to homelessness in that it contains information about people who sought out help, but didn't get referred to one of Compass's programs. An analysis of these households, in conjunction with discussions with the Compass itself, might reveal insights into who gets turned away and why.

A

Appendix

A.1 Description of Datasets

Dataset	Columns
Income Data	• Household ID • Program Name • Sequence • Assessment Date • Total Cash Income • Total Non-Cash Benefits • Total Monthly Income: Sum of cash and non-cash income
Client Date	• Days in Program • Open Date • Exit Date • Exit Reason • Program Status • Age • Client: Race/ Ethnicity • Family Size • Client: Gender • 12 Month Housing Status • 6 Month Housing Status • All Programs • Subsidy Start Date • Household ID • Program Name • Client ID • Case ID

Dataset	Columns

Assessment Data	• Household ID
	• Program Name
	• Assessment Date
	• Assessment Fields
	  – Food
	  – Parenting Skills
	  – Relationships/Domestic Violence
	  – Physical Health and Disabilities
	  – Mental Health
	  – English Language Skills
	  – Child Protective Services
	  – Credit
	  – Child Well-Being
	  – Employment
	  – Child Education
	  – Childcare
	  – Income
	  – Legal
	  – Access to Health Services
	  – Adult Education/Training
	  – Transportation
	  – Substance Abuse
	• ACE Score: ACE is an assessment that measures childhood trauma. Scores range from 0-10. A 10 indicates signficant childhood trauma
	• ACE Score Date: Date of ACE Score Assessment
	• Sequence

Dataset	Columns
Housing Data	• Housing Tracker: Housing Tracker # • Housing Stage • City • Living Situation • Start • Date • Subsidy/ Housing Type • Household ID Program Name

A.2 Results from Exit Score Prediction on Imputed Dataset

Positive Exit	Negative Exit
• 'Left for housing opportunity before completing program', • 'Moved in with family', • 'Increased income and can afford current unit', • 'Moved in to affordable housing', • 'Completed program', • 'Voluntary exit with housing option', • 'Found other housing option', • 'Found permanent housing'	• 'Reached max time allowed without stable housing', • 'Non-compliance with program', • 'Does not meet financial eligibility', • 'Did not comply with Program Rules', • 'Voluntary exit with no housing option', • 'Not making timely progress towards increasing income', • 'Denial of Service', • 'Client Refused Placement', • 'Timed out of housing search', • 'Not accepted', • 'Evicted / Asked to leave unit', • 'Unknown/Disappeared'

A.3 Results from Exit Score Prediction on Imputed Dataset

Compass Clara House

	Mean Imputed Train	Mean Imputed Test	Reg. Imputed Train	Reg. Imputed Test	MICE Imputed Train	MICE Imputed Test
Logistic Regression	0.950000	0.916667	0.950000	0.850000	0.928571	0.916667
LDA	0.978571	0.600000	1.000000	0.633333	0.950000	0.900000
QDA	0.607143	0.583333	1.000000	0.916667	0.985714	0.883333
GNB	0.850000	0.766667	0.935714	0.883333	0.757143	0.666667
MNB	0.728571	0.733333	0.828571	0.883333		
AdaBoost	0.950000	0.916667	0.950000	0.866667	0.950000	0.916667

Compass SF Home

	Mean Imputed Train	Mean Imputed Test	Reg. Imputed Train	Reg. Imputed Test	MICE Imputed Train	MICE Imputed Test
Logistic Regression	0.691589	0.604255	0.786916	0.672340	0.575701	0.561702
LDA	0.697196	0.587234	0.785047	0.663830	0.616822	0.540426
QDA	0.642991	0.612766	0.809346	0.600000	0.639252	0.536170
GNB	0.657944	0.634043	0.757009	0.646809	0.622430	0.570213
MNB	0.607477	0.459574	0.564486	0.502128	0.538318	0.557447
AdaBoost	0.596262	0.591489	0.751402	0.638298	0.594393	0.565957
Polynomial AdaBoost	0.732710	0.685106				
Polynomial Logistic Regression, C=0.5	0.820561	0.706383				
LDA with lsqr and shrinkage	0.777570	0.702128				

A.4 Coefficients for the Return Model

Parameter	Coefficient
CCS	0.2922629444404822
English Language Skills	-0.27363136183592 64
Child Education	0.23616182776799427
Substance Abuse	-0.22191260297876 66
Asian or Asian American'	0.21315219632399796
Adult Education/Training	-0.18900835132432603
White	-0.18624279858898943
ACCESS	0.18500480693716867
Hispanic/Latinx	0.18440512897508873
CFS	0.18158920331727868
Legal	0.15599783851096666
Employment	0.12689627180515528
Physical Health and Disabilities	0.10934771888069575

References

[Com] Compass family services 2017–2018 annual report. Available at https://static1.squarespace.com/static/59c525f71f318df0b9502323/t/5cb0bb0f085229eee3199854/1555086114187/Final+Compass+2017-18+Annual+Report.pdf.

[2] Early, D. (2004). The determinants of homelessness and the targeting of housing assistance. *Journal of Urban Economics*, 55, 195–214.

[3] Hong, B., Malik, A., Lundquist, J., Bellach, I., & Kontokosta, C. E. (2018). Applications of machine learning methods to predict readmission and length-of-stay for homeless families: The case of win shelters in new york city. *Journal of Technology in Human Services*, 36(1), 89–104.

[4] Kube, A., Das, S., & Fowler, P. J. (2019). Allocating interventions based on predicted outcomes: A case study on homelessness services. In *AAAI*.

[5] Melissa J. Azur, Elizabeth A. Stuart, C. F. & Leaf, P. J. (2011). Multiple imputation by chained equations: what is it and how does it work? *International Journal of Methods in Psychiatric Research*, 20, 40–49.

[6] Research, A. S. (2019). *San Francisco Homeless Count Survey Comprehensive Report*. Technical report.

[7] Seabold, S. & Perktold, J. (2010). statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*.

[8] Toros, H., Flaming, D., & Burns, P. (2019). Early intervention to prevent persistent homelessness: Predictive models for identifying unemployed workers and young adults who become persistently homeless. *SSRN Electronic Journal*.