Predicting Women, Infant, and Children Program Participant Churn

Leonard Littleton

Applied Data Science Master’s Program

Shiley Marcos School of Engineering / University of San Diego  
 leonardlittleton@sandiego.edu

Andrew Zazueta

Applied Data Science Master’s Program

Shiley Marcos School of Engineering / University of San Diego  
 azazueta@sandiego.edu

**Abstract**

Woman, Infant, and Children (WIC) is a government assistance program which aids lower income families get the nutrition they need. Due to the program’s importance in helping millions across the United States, they wish to lower their participant churn rate to ensure as many families as possible are benefiting from the program and they are reaching everyone who needs their support. To achieve this goal, machine learning models were used to help predict when a customer might stop using the program. Once WIC provided data on their current participants from South Dakota, feature engineering and data cleaning was performed so the data would be optimal for five unique machine learning models: Naïve Bayes, XGBoost, Logistic Regression, Neural Net, and Random Forest. It was found over 90% of participants in the data were no longer using WIC, meaning the data was highly imbalanced. This was accounted for with synthetic minority oversampling technique (SMOTE) and random under sampling to produce an equal number of both churners and non-churners in the training data set. Choosing the best model was done by examining accuracy and the area under curve (AUC). Random Forests performed the best with an accuracy of 93% and AUC of 0.95. This level of accuracy was achieved after multiple rounds of hyperparameter tuning to ensure the most optimal parameters were selected. Once the model is in deployment, the plan is to contact those who are predicted to churn with information campaigns on how they can best utilize the program to lower the chance of their departure from the program.

**Keywords**

*Keywords*: churn, imbalance, under sampling, over sampling, Random Forest, accuracy, area under curve

**Introduction**

Predicting churn is a common problem almost all data scientists have had to solve at least once in their career or schooling. Its importance is widespread across almost all companies due to the necessity of having high customer retention. In some instances, having high churn not only affects the company, but also the customer. Woman, infants, and children (WIC) is a program that provides low-income families the necessary nutrients and resources they need to remain healthy. When families stop using the WIC program, this can directly result in children not getting their necessities met. Families at-risk of leaving the WIC program can be targeted by building machine learning models that use information of those who have stopped using WIC and those who are still with WIC as training. Once families are identified to be potential churners, information campaigns can be sent out to better increase the chances of keeping them in the program.

**Background**

WIC has been aiding families since 1974 (Owen & Owen, 2003) providing food assistance to women and children. The program is designed for low-income, nutritional high-risk breastfeeding, non-breastfeeding postpartum women, infants, and children aged 1-5. Potential participants travel to WIC clinics to get certified to start the program. If they meet the requirements, they are allotted benefits from month-to-month for the moms and their babies/children. However, participation and recruitment have always been a problem for the WIC program.

Whaley et al. (2020) investigated reasons why participants do not continue their benefits. They found infants in households with incomes less than 100% of the federal poverty level were twice as likely to continue participation through two years old than families above that threshold. Breastfeeding also played a major role in whether families would continue their participation. Infants who were breastfed or partially breastfed for six months or longer were more likely to continue than non-breastfed infants. Lastly, they found families that learned something from the resources WIC offers were also more likely to continue participation through two years of age. In addition to these results, Whaley et al. also found only 77% of participation continued after an infant turned one year old.

The WIC program benefits from their participants involvement. More participation equates to more grants approved each year. The more grants that are approved, the more families get served through the program. Because of this, it becomes mutually beneficial to both the program and the participants to engage fully.

***Problem Identification and Motivation***

The main issue needing to be solved is retaining families in the WIC program so they can continue to receive the resources they deserve. One issue causing churn can be a lack of knowledge on what WIC provides. WIC might be responsible for this issue by not be reaching out enough to families with information on how they can best utilize their program. This issue can be solved with information campaigns designed to let participants know WIC is willing to help them receive all the nutritional benefits they can get. This information campaign will be triggered when a model predicts that a family is at-risk of churning.

The motivation behind finding at-risk families is multi-layered. The main motivation is to help low-income families insure they have less to worry about regarding their health. From a data science standpoint, research into churn model building with health-related data can greatly help a wide variety of companies in the medical field. Features like distance to clinics, income, and family size are all important to the health industry, and data preprocessing methods and model hyperparameter tuning used in this research can be used in multiple projects in a variety of disciplines. Creating a healthier society mutually benefits all, not only those in the WIC program. By helping families remain healthy, a happier community is created.

***Definition of Objectives***

The main objective of this research is to utilize machine learning techniques to predict participants likely to discontinue participating in WIC. The models employed will identify these participants and an information campaign can be triggered to help facilitate more participation. Additionally, feature importance will be considered to help identify the leading reasons why participants do not continue participation. Once these reasons are identified, intervention can be planned to help mitigate those issues.

A secondary objective is to analyze geographic areas and determine which areas are most prone to food insecurity. Insight into this can lead to pop-up clinics or approving more retailers for WIC to also help increase the recruitment and participation. Along with this, analysis into underserved areas can also be useful to continue working toward the goal of recruitment and participation.

**Literature Review**

In this section, three articles relating to WIC and family instability are reviewed along with three articles relating to solving churn from a data science viewpoint. This mixture of journal article topics covers different aspects this project will encounter, whether that be from a business context or a data science context.

***The Impact of Aging Out of WIC on Food Security in Households with Children***

The Special Supplemental Nutritional Program for WIC is a program that assists with food insecurity for children ages 1-5. The cutoff for children to be eligible is 61 months of age. Arteaga et al. (2016) explored the impact of the time period just before and just after this cutoff. They used statistical and empirical approaches to determine if food insecurity increase in the 30 days immediately following the child becoming ineligible. They broke down their data into three main samples: Overall, late starting school, and early starting school. They did find that food insecurity increased, especially with the group of late starting school. The idea here is that these children would be enrolled in kindergarten and benefit from the programs available there. Early starting school children did not experience as significant of a food insecurity. Their research only focused on children who become ineligible due to age. This research is interested not only in participants who become ineligible, but also those who do not renew benefits while still being eligible.

***Family Instability and Material Hardship: Results from the 2008 Survey of Income and Program Participation***

Economic hardships exist inside and out of the WIC program. Heflin (2016) investigated the effects of various instabilities (employment shocks, household formation shocks, residential changes, income changes, household size changes, and disability shocks) on different economic hardships (medical, food, housing, and essential expense). The food hardship is relevant to this research and this research showed that an increase in monthly income contributed to a 2.58% decrease in food instability. The only other factor that led to changes in food stability was having a person with a disability in the household. This led to a 4.33% increase in food instability. This gives two concrete factors that affect food instability. The author studied a few factors across many different instabilities where the focus here is on more factors that contribute to food insecurity alone.

***Can a Better Understanding of WIC Customer Experiences Increase Benefit Redemption and Help Control Program Food Costs?***

Payne et al. (2017) examined customer experiences that could lead to higher WIC redemptions and lower food cost. They held a series of focus groups and learned from WIC participants factors that contributed to poor experiences, which in turn could have led to lower participation. The four main factors they revealed were: retailer size, store crowding, employee-customer interactions, and retailers’ WIC program management. All four factors played pivotal roles in discouraging WIC participants from participating fully. Their research was limited to convenience samples in one state (New Mexico). These findings are good but might not be representative of the entire WIC participant population. This research looks to dig into more of the reasons that might lead to a participant not participating.

***Deep Learning in Customer Churn Prediction: Unsupervised Feature Learning on Abstract Company Independent Feature Vectors***

Spanoudes and Nguyen (2017) outlined how they built deep neural networks to target at risk customers of leaving a variety of companies to increase customer retention. To do this, the team researched deep feed-forward architectures with “logistic regression using a single artificial neuron up to the implementation of a Multilayer Perceptron” (p. 14). From their studies, they found that this architecture would be the one they would build their model upon. They created four layers to this network: a linear hidden layer, two dropout layers, and an output layer that used a softmax layer architecture. The effectiveness of their model was decent with an average accuracy of about 75%. From this paper, feed forward neural networks show promising prediction power towards an array of unique use cases in predicting customer churn.

***An Effective Classifier for Predicting Churn in Telecommunication***

Pamina et al. (2019) decided the best way to predict churn was by using ensembling models, as well as KNN. The two ensembling models created were Random Forest and XGBoost. Out of the three, XGBoost performed the best with an accuracy of 79% with Random Forest at 78% and KNN at 75%. With these models, they also found certain predictors which contributed most to churn, which is important for businesses to help fix their customer retention issues. The data set used in their research was imbalanced with only 27% of the samples being churned versus 73% non-churned.

***Comparison of Two Main Approaches for Handling Imbalanced Data in Churn Prediction Problem***

This leads to studying Nguyen and Duong’s (2021) work, which highlights how to tackle class imbalance when preparing the data for modeling churn. In their paper, the team outlined two main strategies to handling imbalanced classes: resampling and “cost-sensitive learning methods that adjust the relative cost of errors during model training” (Nguyen & Duong, 2021, p. 29). The two resampling methods chosen were SMOTE and Deep Belief Network (DBN) and the two cost-sensitive learning methods were focal loss and weighted loss. Logistic Regression and XGBoost were the models chosen to test which data wrangling methods yielded the highest accuracy. The cost sensitive learning methods along with the XGBoost model performed the best with an area under curve (AUC) of 0.9115 (weighted loss) and 0.8925 (focal loss). However, both resampling methods greatly increased the accuracies of both models. For example, the baseline logistic regression model without SMOTE or DBN had an AUC of 0.5759. Resampling with SMOTE increased the AUC to 0.7568 and DBN increased the AUC to 0.6431, both far better than the original performance. Using the XGBoost machine learning algorithm along with focal and weighted loss is a solid strategy to handling imbalanced churn data, but it does require more computation power compared to their resampling counter parts. If the data were to contain millions of records, resampling might be the better choice. If the data is smaller, then cost-sensitive learning methods might be the best option due to it performing better in this study.

**Methodology**

The methodology followed in this project includes an end-to-end data science pipeline. The first step was data collection. When the data was collected, exploratory data analysis was applied to the data. This allowed the familiarity with the data to be gained. After the data was thoroughly explored, data cleaning techniques were then applied to the data. This included handling missing values and outliers, determining useful features, dropping un-useful features, and identifying a target column. In addition to the steps above, binary features were also converted to 0s and 1s. This allowed machine learning models to handle that kind of data better. Categorical features were also converted to dummy variables.

Once the data was cleaned, the next step performed was feature engineering. Location data is often difficult to analyze especially without exact addresses. New features were created to utilize some location data that could be useful. Date columns tend to not be useful in prediction problems. The target feature of this data set was engineered based on a few date columns.

The last part of the Exploratory Data Analysis and Data Preparation step was to perform some correlation analysis on the features. When two or more variables are correlated, that leads to collinearity. It is important to account for this. The relationships between the predictor features and target feature are also examined.

***Data Acquisition and Aggregation***

Custom Data Processing (CDP) and the state of South Dakota supplied the data for this research. CDP is a software solution company that hosts the data warehouses for many of the WIC agencies, including South Dakota. Permission was requested and granted to utilize this data. HIPAA considerations were needed to utilize this data since it contained PII data for real-life participants of the WIC program.

The data warehouse follows a star schema where data is stored in dimension and fact tables. This data is sent in from each individual agency. Prior to obtaining the data for research purposes, it needed to be deidentified. This was performed using SQL and removing any features that contained PII data: names, addresses, and phone numbers for example. Once the data was approved, SQL was used again to join the various dimension tables to the fact table of interest. This was exported to a csv and was how the dataset was obtained. Data in this fact table is updated monthly. For the purposes of this research, stale data is being used. However, in practice, new data would be added and analyzed in real-time.

Python and Jupyter Notebooks were the chosen coding platform to perform analyses. Python is a powerful scripting language popular with data science techniques. It has multiple, useful libraries that can be leveraged for exploratory data analysis (EDA), data preparation, modeling, and evaluation.

***Exploratory Data Analysis***

It is important to explore the data to get a feel for data distributions, feature values, missing data and outliers, and feature relationships. Three types of features were considered in this analysis: binary predictors, non-binary predictors, and binary target.

The binary predictors consisted of 25 features. Each of these features was examined for frequency distributions. Many of these predictors were heavily imbalanced. This could lead to models favoring these predictors due to the imbalance. It is something that should be monitored and evaluated during the evaluation phase of modeling.

The non-binary predictors consisted of three features: household size, household income, and poverty level. The analysis showed that most families fell into a middle tier for household size. Household income was most represented on the lower end of the spectrum. This supports the program being designed for lower income families. Finally, the poverty level also followed the income trend, which holds with the expectation.

The target column, certification flag, is an engineered column based on certification start date and certification end date. If the current date falls between those two dates, the flag is set to 1. Otherwise, it is set to 0. The analysis of this feature shows a class imbalance in favor of not certified. Analysis was also performed on the non-binary features in relation to the target feature. This continued to follow the expectation. Lower household incomes were more likely to be certified. Poverty level showed a somewhat even distribution for those who were certified. Certified individuals also seemed to fall in the midrange of household sizes.

***Data Quality***

The data set in total had 698,729 rows and 97 features. Some features included keys which are meant for identification of the participant or agency, which are not important for model building. Of these keys, the participant key was left to be able to identify the potential churner for the ability to contact them. One important thing to note about the data was that each row did not belong to a unique participant; only 139,359 unique participants existed in the data set, with many having more than one row. This duplication of participants was most likely a result from the joining of data frames prior to data preprocessing, which resulted in multiple features having missing values. Of the 97 columns, 17 had to be removed due to having less than 70% of the rows non-null.

***Data Quality Issues.***

Data quality issues centered around outliers and categorical features. Only three columns in the data set which would be used for modeling had continuous numerical data: household income, household size, and poverty level. The outliers were identified by creating boxplots for each feature (see Figure 1) and the rows which had these outliers were removed. However, not all outliers for household income were removed. This was because the data was accurate and a natural occurrence, which could happen again in future data sets, so the final model would need to be trained for this occurrence. The remaining steps needed to preprocess the data for modeling solved the remaining data quality issues, which were mostly centered around how to handle categorical features and correlated columns.

**Figure 1**

*Boxplots of Continuous Features in Data Set*

Chart, box and whisker chart

Description automatically generated

*Note*. Plots were made using Python’s Seaborn library. The top left plot shows household income (measured in USD), the top right plot shows the household size, and the bottom plot shows the poverty level (derived from “[dividing] income by the poverty guideline and [multiplying] by 100” (Hayes, 2022)).

**Feature Engineering**

Features which only took two values were engineered by changing one of the values to 1 and the other to 0. Some categorical features had more than two values, like primary language. This was converted to be 1 for *English* and 0 for other languages present in the data set. However, other columns which had more than two categorical values were changed into *n-1* dummy features, with *n* being the amount of unique categorical values found in a column. For example, mother’s education level took the values below high school, college, grad or GED equivalent, unknown, not applicable, or master’s or higher. This feature changed into five binary features, eliminating below high school to avoid correlation between the newly created columns. Correlated features were identified by the creation of correlation matrices. If two columns had a correlation of 0.7 or above, one would be removed.

Handling the three non-categorical numeric columns mentioned earlier was different as it could not be binned into 0 or 1. These were normalized so that the max value they took was 1 with the minimum value being 0. This avoids having different units of measurement affect the model’s performance, as household income contains much greater values than household size, for example.

Another area where new columns were created was determining if a clinic existed in the same city as the household. If it did, the new column called ‘same city?,’ was marked 1, and 0 if not. This avoided dealing with exact participant addresses to calculate distance between the home and clinic, and rather create a binary feature which models would have an easier time handling. This same step was repeated for the participants local agency city and household city.

Lastly, the target column had to be created from the values of two other columns: certification start date and certification end date. If the current date was greater or equal to the certification start date and less than or equal to the certification end date, then the participant was certified (marked 1). Else, they were not (marked 0). This resulted in a target feature which was imbalanced, with the majority class being churned participants (~91%). In production, this feature would be dynamic due to the current date changing day to day.

**Modeling**

Before modeling could begin, the data needed to be separated into three sets: training, validation, and testing. Each data set was partitioned to have an identical ratio of majority and minority class to insure no one set received too many or too little of the minority class instances. 70% of the data went into the training set and the remaining 30% of the data was split evenly between the validation and test set.

To account for imbalanced data, synthetic minority oversampling technique (SMOTE) and under sampling was used to ensure the majority and minority class were balanced. Both resampling techniques were only used on the training data set. SMOTE helps even the ratio between the classes by generating new minority data points rather than duplicating. The SMOTE function came from the Python library imblearn, and the sampling strategy argument was set to 0.1. This transformed the ratio of minority class to majority class to 0.1 rather than 0.09 which it was before using SMOTE, a minor change. After this step, random under sampling was used to decrease the majority class. The sampling strategy argument was set to 1 to create an equal number of both classes. After these transformations, the training set had a total of 88,188 instances.

A total of five models were made: Naïve Bayes, XGBoost, Random Forests, Logistic Regression, and Neural Net. Naïve Bayes acted as a baseline model to measure the other model’s performances. Ideally, each model will perform better than Naïve Bayes. Since the data set is imbalanced at 91% of the participants labeled as churned, the ideal accuracy for the models would be above 91%. Of the five models, Random Forests performed the best with a 93% accuracy, beating the baseline accuracy.

***Naïve Bayes***

Naïve Bayes performed the worst of the five models as expected. Using every feature in the data set for modeling, Naïve Bayes was only able to achieve a 61% accuracy. Naive Bayes did well at predicting the majority class with an f1-score of 0.73 but performed poorly predicting the minority class with an f1-score of 0.30. No hyperparameters were tuned to get a true minimal performance to beat.

***XGBoost***

The first non-baseline model built was XGBoost. The initial XGBoost model was fit to the training set to find the top 20 most important features. This step was important to save time for the hyperparameter tuning. If all 88 features were used for model tuning, it was estimated that model training would take over two hours. The initial XGBoost model with no tuning yielded an 88% accuracy on the validation set, which was already much better than Naïve Bayes.

After the 20 most important features were found (see Figure 2), a new training set containing only those features was created. This data set was to be used for the tuning steps listed next. This step is not only important for modeling, but also important for WIC. This allows WIC to identify which features contribute most to predicting if a customer would churn or not. From the list, terminated count and disqualified count are the two features which aided the models’ predictions most. WIC can monitor the data from these columns to get a better idea how to identify a potential churner.

**Figure 2**

*20 Most Important Features for XGBoost*

Chart

Description automatically generated

*Note:* This list of the 20 most important features for XGBoost was found before model tuning. These features were then used to tune the XGBoost hyperparameters to save training time.

The grid search function from sklearn allowed the models to use a range of hyperparameter values with the desired intervals in between the maximum and minimum values assigned. The first hyperparameters that were tuned were *max\_depth* and *min\_child\_weight*. The original model had a *max\_depth* of five and *min\_child\_weight* of one. After tuning, the best *max\_depth* was found to be nine, while *min\_child\_weight* remained optimally set at one. The next hyperparameter tuning was for gamma and included the new max depth value. After tuning was completed, a gamma value of 0.4 (originally 0) was found to be best. The last hyperparameter tuning conducted was for the arguments *colsample\_bytree* and *subsample*. Initially, both these values were set to 0.8, but after tuning was completed, both values were changed to 0.9.

After these steps, the model was ready predict the target column classification flag in the validation and test sets. The original training set with all features was used. Unfortunately, the tuning only increased the accuracy of the validation set to 89%, a 1% increase. The test set had an accuracy of 89% as well.

***Logistic Regression***

The next series of models were ran using the default parameters and the best performing model of this series was evaluated to get the best model. The first model of this series tested was Logistic Regression. Logistic Regression is a great choice for a classification model because of its simple nature. The model was built using default parameters. It yielded a high accuracy of 89%. This was a good result over the baseline accuracy. This result was not the highest of this series, however.

***Neural Net***

The second model in this series was Neural Nets. These models are great at learning complicated patterns within the data. The interpretability is not high though. The Neural Net model yielded an accuracy of 89%, which was slightly higher than the Logistic Regression model. This is another good result over the baseline accuracy. This model finished second in this series of modeling.

***Random Forest***

The last model considered in this series of models was Random Forest. Random Forest is an ensemble method that performs multiple decision trees at the same time. It is another model that is easy to interpret and implement. The results of Random Forest were astounding. The default parameter model yielded an accuracy of 92.37%, which was the highest of any model tested. This model was investigated further to see how high the accuracy could go.

Random Forests have a variety of parameters that can be tuned. Using intuition and domain knowledge, the parameters that were tuned were *n\_estimators, max\_features, max\_depth, min\_samples\_split, min\_samples\_leaf, criterion, bootstrap, and oob\_score.* The randomized search algorithm was used to get in the vicinity of the best parameters. This was ran using 100 iterations and a 5-fold cross validation. The best parameters using this approach were: 400, sqrt, 40, 2, 1, gini, false, and false, respectively. Using the best parameters found by the random search algorithm, the accuracy increased to 92.91%, which was an increase of a 0.5% to the default parameters. This increase is negligible, but it could be substantial with this volume of data.

In addition to the random search, a grid search was also performed to confirm whether the random search found the best parameters overall or not. The grid search returned all the same parameters except n\_estimators = 300. The accuracy using this parameter was worse than the original random search parameters, so the conclusion was the random search found the best parameters.

**Results and Findings**

Random Forests performed the best with a 93% accuracy. Another method which was used to evaluate model performance was the AUC. The higher an AUC, the better the model. XGBoost had an AUC of 0.93, while the Random Forest model had an AUC of 0.95. Random Forest was the best performing model of all the models tested and is the model chosen to move forward with. Finally, Random Forest also had the best F1 score. This metric is a combination of precision and recall so that both can be compared at the same time. In the majority class, it yielded a score of 0.96 and in the minority class, it yielded a score of 0.70, which was best in both cases for all models tested.

It was surprising to see several models perform well on the data. This gave the flexibility to fine-tune the model of choice rather than be forced to pick a model that may not be great for the data. XGBoost and Random Forests both performed well. However, Random Forest edged it out by just a little. This model can now be deployed to identify WIC Participants that may not recertify, or churn. The next step will be to trigger an information campaign for the participants identified by this model to help persuade them to recertify.

**Discussion**

The research presented incudes an end-to-end data science process to predict WIC participants who are likely to churn. The process began with a literature review where similar studies were reviewed. This included studies where food instability is measured when children age out of the WIC program, survey information on family instability, and whether a better understanding of WIC experiences can lead to more participation. There were also studies that looked at multiple approaches using machine learning with customer churn problems: using deep learning for customer churn predictions, an effective classifier for telecommunication churn predictions, and how to handle imbalanced data with churn problems.

The next step included an extensive exploratory data analysis on the data that was provided by CDP and the state of South Dakota. The data was analyzed for missing data, outliers, and data quality. Common techniques like imputation were used to handle missing data. Outlier detection revealed data that was likely bad. Feature engineering was performed to calculate the target column as well as transform binary features into values appropriate for modeling.

The final step in the process included examining several machine learning models to classify the data into two values: certified or not certified. The models investigated included Naïve Bayes, XGBoost, Logistic Regression, Neural Net, and Random Forest. XGBoost and Random Forest both performed well, and both went through hyperparameter tuning to determine the best model. Random Forest was the best performing model yielding an accuracy of 93%. This bested the baseline accuracy of solely predicting the majority class.

The results of this research are significant. The Random Forest model that was trained and hyperparameter tuned will provide valuable predictions that should increase participation in the WIC program. Once the model is deployed, it will identify participants likely not to recertify. This will allow an information campaign to trigger to notify such participants that they should recertify.

Time constraints did limit the results of this study. Other facets about WIC participation were not able to be investigated. This included identifying areas where food insecurity may be the highest. It also included not being able to analyze geographic regions where areas might be underserved. Additionally, some data that was allotted for this project was not able to be used as time would not allow. Due to the sensitive nature of the data, approvals were needed to use the data, which severely cut down the time available to work with the data.

One last limitation is that the data came solely from participants in South Dakota. It is uncertain whether the findings here can be applied to other states. Repeating the research with data nationwide could help uncover insights more appropriate for the entire population of WIC participants.

**Conclusion**

The non-baseline models performed well on this data and produced high accuracies. This result is promising because it exemplifies how well machine learning models performed with data supplied by WIC.

Feature engineering is a pivotal step in exploratory data analysis. The data used here needed a target column to be created. In addition, there were transformations needed for the models to work with the data. With more time, more feature engineering could have been performed to increase the performance of the models.

WIC is a tremendous program that serves many people nationwide. It is important for those eligible to participate to continue their involvement. The model created will help identify participants likely not to recertify and intervention can happen to ensure their participation continues.

**Recommend Next Steps/Future Studies**

It is recommended that the state of South Dakota adopt this model and apply it to their production data. The next step after deployment will be to determine the best form of information campaign. The best suited options include phone calls, emails, or text messages. The system can automatically perform one of these tasks when a participant is identified by the model as someone likely to not recertify. Participation can then be tracked to verify that participation increases as a result of this model.

Future work can analyze other states and perform similar analyses to see if participation can be increased in those states as well. Additionally, more time can be spent to analyze the features that time did not allow for in this research. This includes investigating areas where food insecurity is the highest. Another application that can be studied is identifying underserved areas.

**Acknowledgments**

We would like to thank the state of South Dakota and Custom Data Processing for providing the data used for this project. Additionally, we would like to thank the University of San Diego for the opportunity to experience working on this end-to-end data science project. So much has been learned from the Applied Data Science program and we both look forward to applying these skills to future problems we encounter.

**References**

Arteaga, I., Heflin, C., & Gable, S. (2016). The impact of aging out of WIC on food security in households with children. Children and Youth Services Review, 69, 82–96. <https://doi.org/10.1016/j.childyouth.2016.07.015>

Hayes, A. (2022, May 25). *Federal Poverty Level (FPL).* Investopedia. https://www.investopedia.com/terms/f/fpl.asp

Heflin, C. (2016). Family instability and material hardship: Results from the 2008 survey of income and program participation. Journal of Family and Economic Issues, 37(3), 359–372. https://doi.org/10.1007/s10834-016-9503-6

Nguyen, N., & Duong, A. (2021). Comparison of Two Main Approaches for Handling Imbalanced Data in Churn Prediction Problem. Journal of Advances in Information Technology, 12(1).

Owen, A. L., & Owen, G. M. (2003, April 23). Twenty Years of WIC: A review of some effects of the program. Journal of the American Dietetic Association. https://www.sciencedirect.com/science/article/abs/pii/S0002822397001910

Pamina, J., Raja, J., Bama, S., Soundarya, S., Sruthi, M., Kiruthika, S., Aiswaryadevi, V., Priyanka, G. (2019). An Effective Classifier for Predicting Churn in Telecommunication. Journ of Adv Research in Dynamical & Control Systems, 11(1).

Payne, C. R., Niculescu, M., Guthrie, J. F., & Mancino, L. (2017). Can a better understanding of WIC customer experiences increase benefit redemption and help control program food costs? Journal of Hunger & Environmental Nutrition, 13(2), 143–153. <https://doi.org/10.1080/19320248.2017.1364191>

Shiley, 2022. The Shiley Marcos School of Engineering: USD  *University of San Diego*  DOI: <https://doi.org/12.3456/1234567.1234567>.

Spanoudes, P., & Nguyen, T. (2017). Deep Learning in Customer Churn Prediction: Unsupervised Feature Learning on Abstract Company Independent Feature Vectors. https://doi.org/10.48550/arXiv.1703.03869

Whaley, S. E., Martinez, C. E., Paolicelli, C., Ritchie, L. D., & Weinfield, N. S. (2020). Predictors of WIC participation through 2 years of age. Journal of Nutrition Education and Behavior, 52(7), 672–679. https://doi.org/10.1016/j.jneb.2019.12.015.