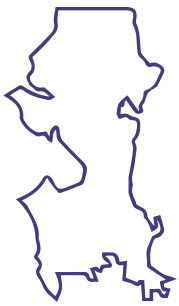

# Exploring + Modeling Seattle's Aging and Disability Services



**Jane Christen + Leo Harth**
Planning by Numbers
Spring 2019

*Course Led By Professor Megan Ryerson + TA Josh Davidson*

Jane Christen + Leonardo de Castro Harth
May 8, 2019
Planning By Numbers Final Term Project
Professor Megan Ryerson + TA Josh Davidson

# Exploring and Modeling Seattle's Aging and Disability Services

*Regression, Predictive Modeling, and Long Term Implications*

## Introduction

In the preceding decade, the Seattle King's County region has encountered a substantial surge in the number of older adults with various service-aid needs and disabilities. Generational specific growth, seen across the United States (U.S.) and particularly in Seattle, is colloquially referred to as the "age wave". This unprecedented wave is anticipated to escalate dramatically as the baby boomer age group continues to mature.[1] [2] Currently, 17% of the regional population is over the age of 60, and this group is expected to grow by more than 25% by 2040.[3] Substantial demographic fluctuation poses numerous challenges such as accurate budget forecasting urgency, prioritization of discretionary funds, compressed long-term hiring planning, potential service alterations, and others. Yet, these shifts also champion new opportunities for partnerships, advocacy, education, healthy aging, and community engagement.

The Aging and Disability Services (ADS) was founded in 1971. Subsequently, in 1973 per the federal Older Americans Act (OAA), the State of Washington apportioned 13 area agencies. This move solidified ADS's Seattle regional oversight. ADS now aids over 40,000 clients each year (duplicated counts linked). [4] Via sub-contracting, ADS provisions a network of in-home and community services and supports for seniors and those with disabilities.These forms of assistance address the physical, financial, and social aspects that impact the health and well-being of older community members. ADS strives to guarantee that all elderly members of society experience solid health and may age in place as desired.

One step in ensuring that the older population age wave is appropriately planned for by Seattle, Washington and similar cities across the U.S. is investigating the highest used services, and then understanding which clients rely on these services. Furthermore, this analysis seeks to identify vulnerable elderly populations with low service use, but theoretically high need, and hypothesize why use counts are low. Understanding usage discrepancies and gaps can highlight regional service distribution equity issues that require addressing. Finally, this data-driven exploration examines predictive modeling capabilities in future Seattle area and potentially U.S.-wide aging budgetary and equity-oriented planning efforts.

---

[1] "Baby Boom Generation," Historical data, United States History, May 2019, https://www.u-s-history.com/pages/h2061.html.
[2] The term "Baby Boom" is used to identify a massive increase in births following World War II. Baby boomers are those people born worldwide between 1946 and 1964.
[3] Maureen Linehan, "Area Plan: Area Agency on Aging Seattle-King County, Washington 2016-2019," Online (Seattle, WA: ADS, October 2015).
[4] Maureen Linehan, "Area Plan: Area Agency on Aging Seattle-King County, Washington 2016-2019," Online (Seattle, WA: ADS, October 2015).

## Data and Methods

### *General Overview*

The primary dataset used for the purposes of analysis was appropriated from the ADS database that captures client-level service provisioning. This data was downloaded specifically from the city of Seattle's open source data portal (https://data.seattle.gov).[5] Each ADS dataset reviewed contained a highly descriptive list of categorical variables delinating which service was utilized by the client. In addition to the categorical descriptor variables (e.g., income, race, region), a variety of detailed binary variables are also included in the data set (e.g., help, disabled, veteran, etc.); see ***Appendix Figure B***. During the data cleaning process the team determined the overall analysis would be strengthened by transforming many of the service use categorical variables into dummy variables (e.g., used_transportation, used_nutrition). One particular constraint of the ADS datasets was that very limited information was available regarding how the client-level data was collected and by whom (e.g., client, service worker, caretaker, family member). This limitation led to some minor speculation regarding response assumptions that are highlighted as necessary throughout the following report.

The ADS client-level data from 2010-2016 was initially considered. Since the team was interested in forecasting aging needs and trends we first cleaned only the 2015 and 2016 dataset years. Upon mapping out the data transformation possibilities, the team decided to explore the potential predictive possibilities such as using 2015 data to prognosticate 2016 outcomes; this is described in further detail in subsequent sections of the paper. This development was later perceived as having a high potential of continuing flaws since many of the same client using services in 2015 would presumably continue using those same services in 2016. After some additional consideration and at the recommendation of University of Pennsylvania Ph.D. candidate, Xiaoxia Dong, 2015 ADS data was cleaned and used to run simulations based on the developed predictive model in order to develop a more holistic understanding of the dependent variables' reactions in relation to multiple independent variables.

### *2015 + 2016 Data Wrangling*

Initially, the 2015 and 2016 ADS datasets contained approximately 300,000+ rows and 45 variable columns. Each row in the dataset represents a service user that utilized one of the listed services. During the data cleaning process, this large information set was culled down to 14,000 observations and 35 variable columns. The removal of roughly a couple hundred thousand observations was regarded as potentially limiting since the team originally thought the dataset was incredibly rich, only to find out that the high presence of incomplete observations were initially present.[6] As this project's intent was to highlight the supports and services received by the aging population, the data age range threshold was set to capture only those ADS clients older than age 65.[7] Finally, the outcome supports provided to the aging clients were summarized, and the service variables reporting the highest use frequency were identified; see ***Figure 1*** for all original areas of service options provided to clients. The four variables per request count and team interest were:

- Case Management (108,827 initial observations),

[5] ADS, "Aging and Disability Services - Client Level Data | City of Seattle Open Data Portal," Seattle Open Data, April 2019, https://data.seattle.gov/Community/Aging-and-Disability-Services-Client-Level-Data-20/bk4b-z4j9.

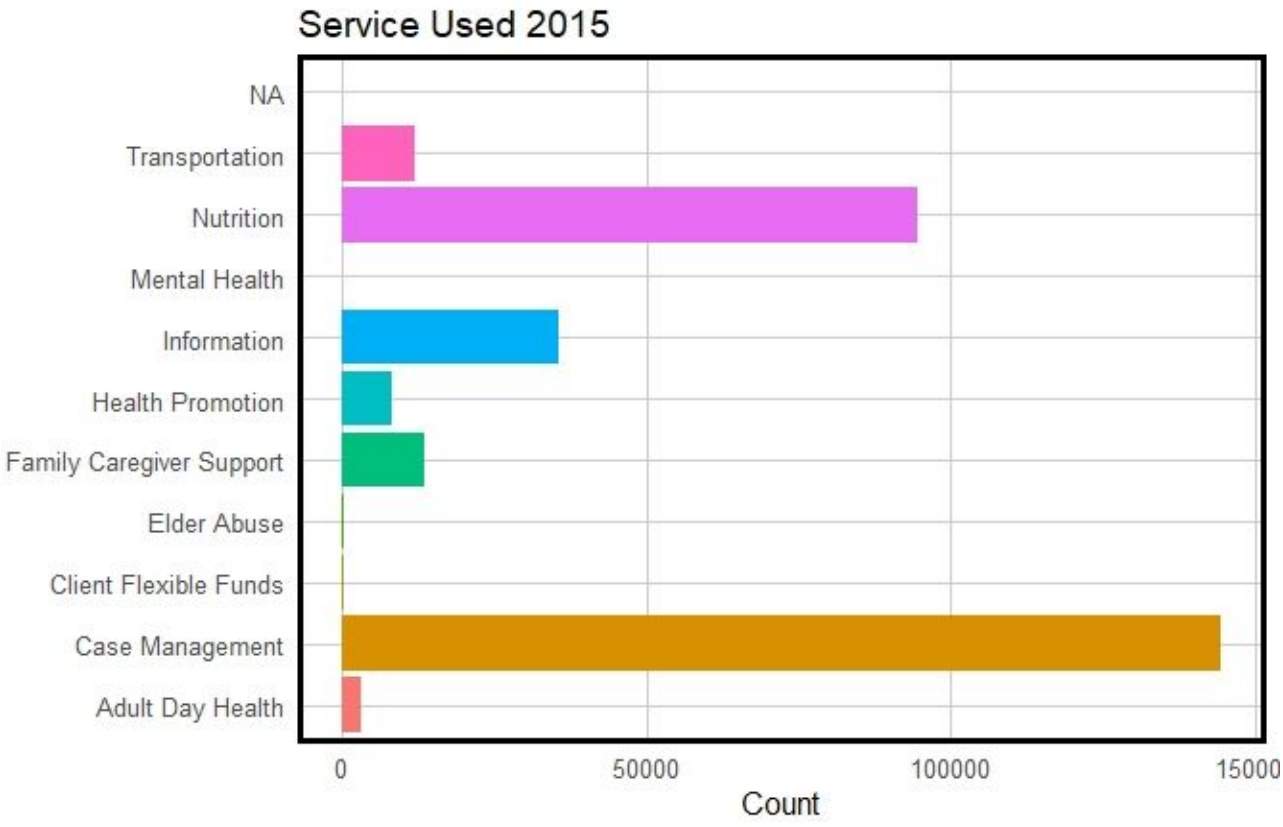
[6] Per the metadata and Area Plan report, some of this information was scrubbed and removed from the dataset to protect client privacy.

[7] The original raw 2015 and 2016 ADS datasets contained client information on children and any Seattle region member that was purported to be disabled an dependant on ADS services.

- Nutrition (78,102 initial observations),
- Information and Assistance (25,923 initial observations), and
- Transportation (9,960 initial observations)

The three top known outcomes of Case Management, Nutrition, and Information and Assistance were adopted as dependent variables of analysis. In addition to these three, Transportation was also included as a dependent variable, as a team's topic of interest. These were investigated in tandem with the 25 independent variables of activities and surveyed service need. The 25 independent variables explored were binary in nature, and additional categorical variables such as income, race, and geographical region were also incorporated in the regression analyses.[8]

***Figure 1 -*** *All Areas of Service Provided to Clients. The four most popular services (Case Management, Nutrition, Information and Assistance, and Transportation) were selected from this series.*




[8] Binary variables are also referred to as dummy, indicator, or design variables

***Service Use Frequency***

The histograms below display the frequency that users of these four services requested those services within the 2015 timeframe. These plots provide some insight regarding the relationship between the nature of the service and the service user. Case Management service users are very bifurcated and utilize the service either once a year (83%), or twelve times per year (16%); see ***Figure 2***. Due to the 12 or one designation, this might indicate that some users rely on this service on a monthly (12-times per annum) to assess the evolution of their cases. Information and Assistance is frequently used once or twice (respectively 74% and 18%); see ***Figure 3***. The use of Transportation highest count peaks in are indicated at two and four times per year (respectively 48% and 21%); see ***Figure 5***. The even number suggests that round trips are double counted. The use of Nutrition services is the most recurrent one: 16% of the users utilize nutrition services more than 50 times per year; see ***Figure 4***.

***Figure 2*** - *Frequency of Use: Case Management*

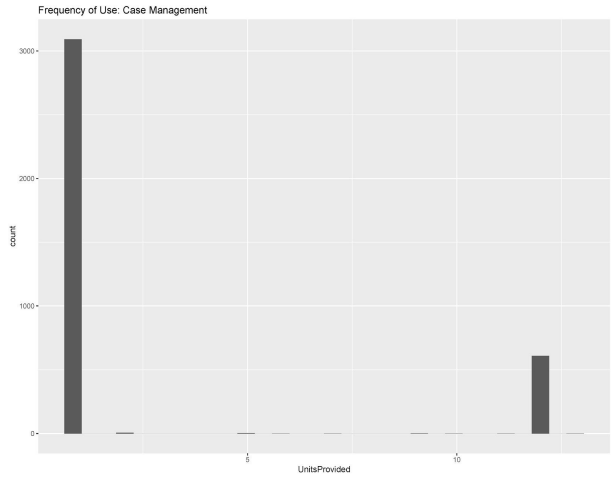



***Figure 3*** - *Frequency of Use: Information & Assistance*

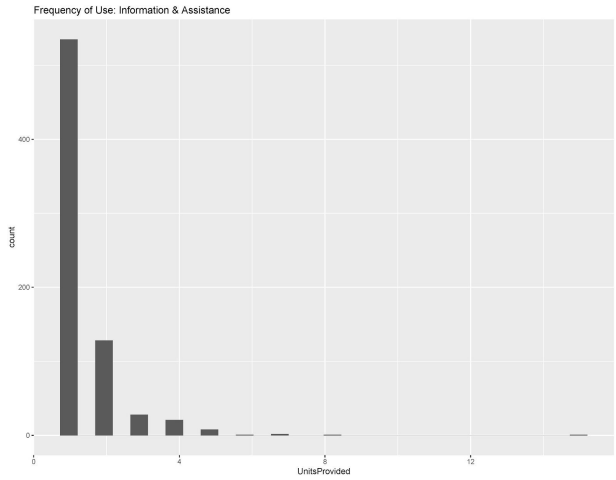



***Figure 4*** - *Frequency of Use: Nutrition*

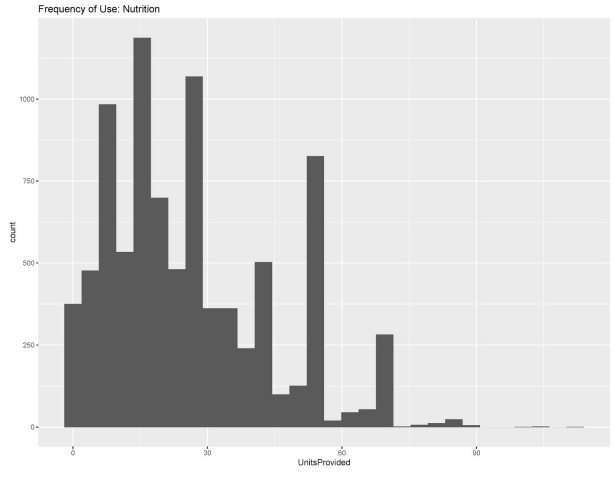
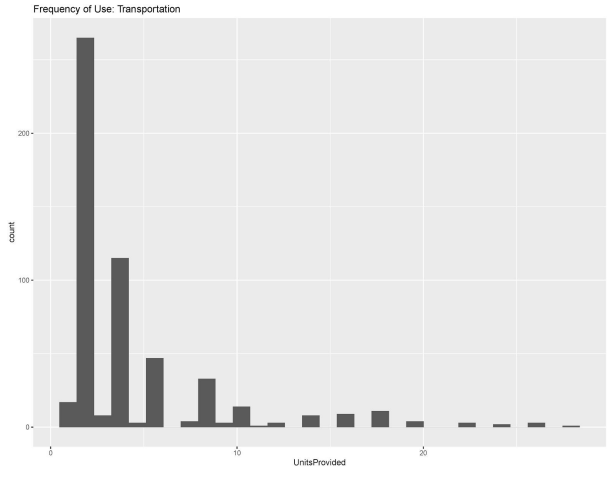



***Figure 5*** - *Frequency of Use: Transportation*



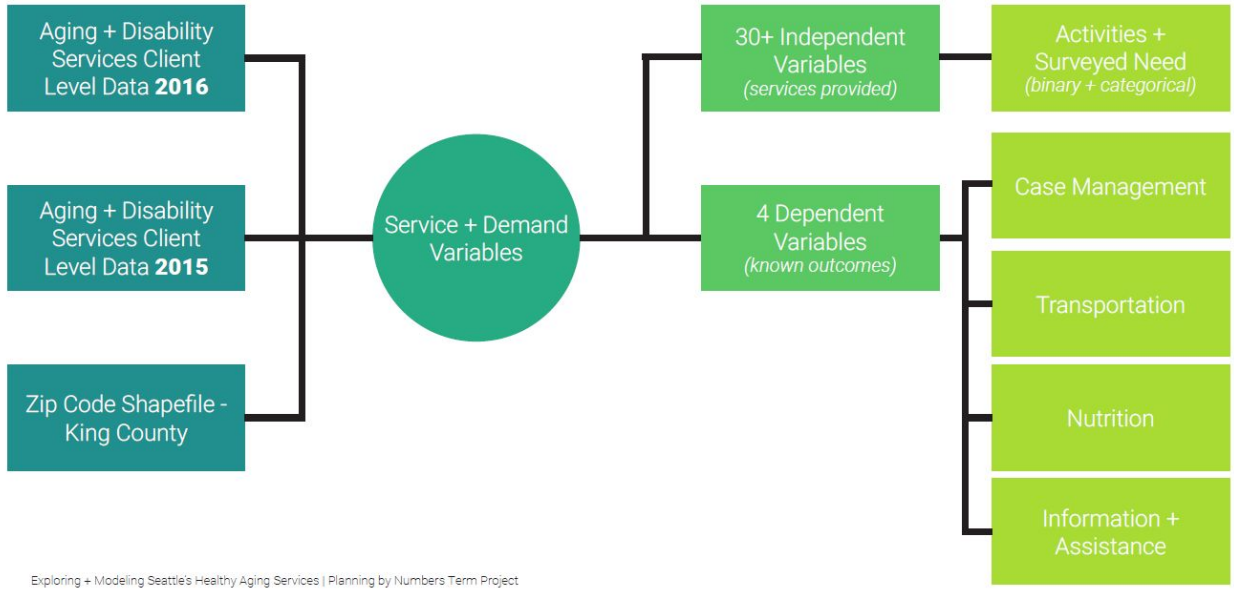
***Continuous Variable Pursuit - Distance***

At first, the team hypothesized that the continuous variable of distance from ADS neighborhood centroids to Seattle's downtown, central business district (CBD) might furnish an additional variable of significance. Therefore, the Seattle neighborhood zip code shapefiles were downloaded and linked to the specific ADS neighborhood designations as reported in the metadata sheet. The centroids of each neighborhood was then calculated. Finally, the distance from each respective neighborhood's centroid to the CBD was measured in miles. Ultimately, when the continuous variable of distance was incorporated into the developed model, high p-values underscored the low level of significance and collinearity. This depressed influence is likely due to collinearity with a fixed effect from the overarching region. Yet, in some of the constructed models, the variable "Distance_Miles" presented elevated levels of significance. Therefore, distance was deemed significant in some cases, and in others the neighborhood fixed effect was more influential.

***Categorical Variable Recoding - Service Use***

The integer identifying the type of service provided to a client was initially coded as a categorical variable. For example, if "8" was listed, this meant the client used "Transportation", and if "11" was listed, then the client had utilized the "Nutrition" service. Since the team's intent was to produce a logistic regression outcome, these respective service use categorical variables were recoded into dummy variables (e.g., did used, or did not use).

***Figure 6*** *- A flowchart depicting the data collection, cleaning, and joining process that aggregated to form the final datasets that were then manipulated via regression and predictive modeling.*



## Data Exploration

### *Understanding Service Provisionment*

Once the data was cleaned, the team began exploring the data developing an understanding of the emergent service use patterns. First, the frequency of service use by the four dependent variables was visualized; see ***Figure 7***. Here, it is important to note that although Case Management is the overall most requested service, in the clean dataset, nutrition is the most used. The Nutrition count makes up 64% (8,781), Case Management 27% (3,716), Information and Assistance 5.2% (725), and Transportation 4.0% (554). To understand the geographic disbursement of the service use frequency, the same use factors were observed by area; see ***Figure 8***, ***Figure 9, Figure 10***.[9] Per the ADS service catchment neighborhood designations, the South Urban, North Urban, SE Seattle, and Downtown area present the highest service frequencies.

***Figure 7*** - *Frequency of Service Use Overall by Dependent Variable*

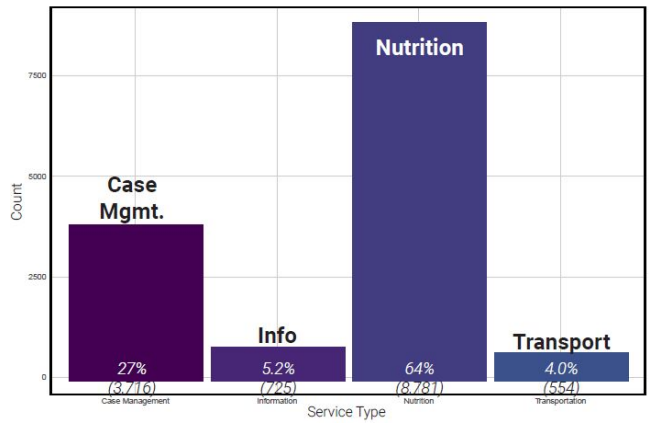
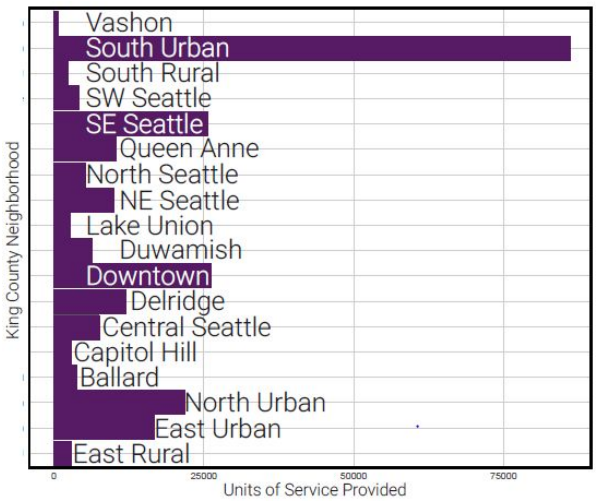



***Figure 8*** - *Frequency of Service Use by Area*




[9] These stats are relative to the cleaned dataset of 14,000 observations.

***Figure 9*** *- ADS Area of Analysis by Seattle's King County Area/Neighborhood*

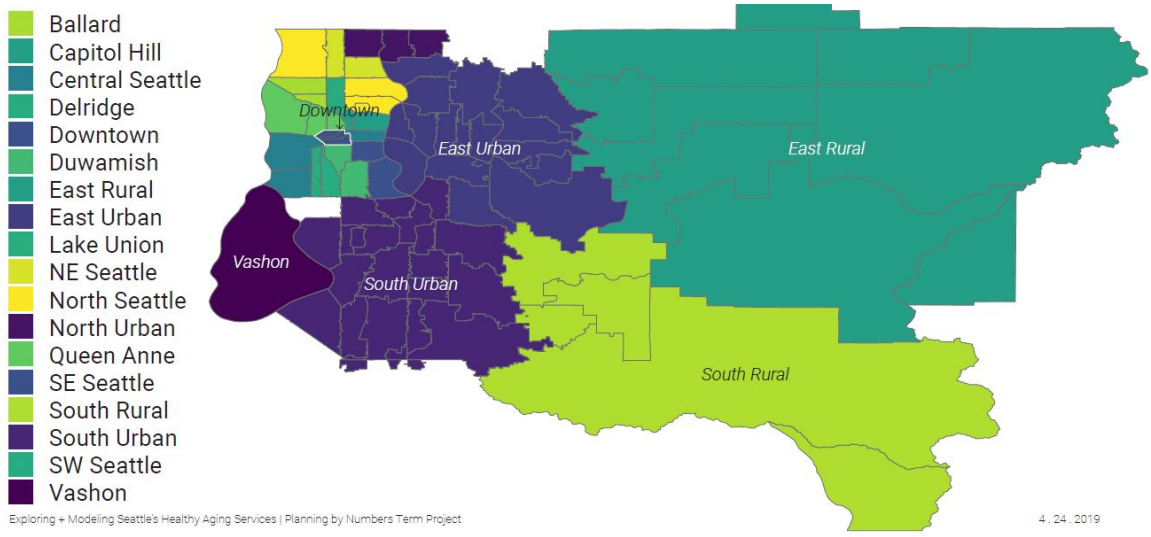
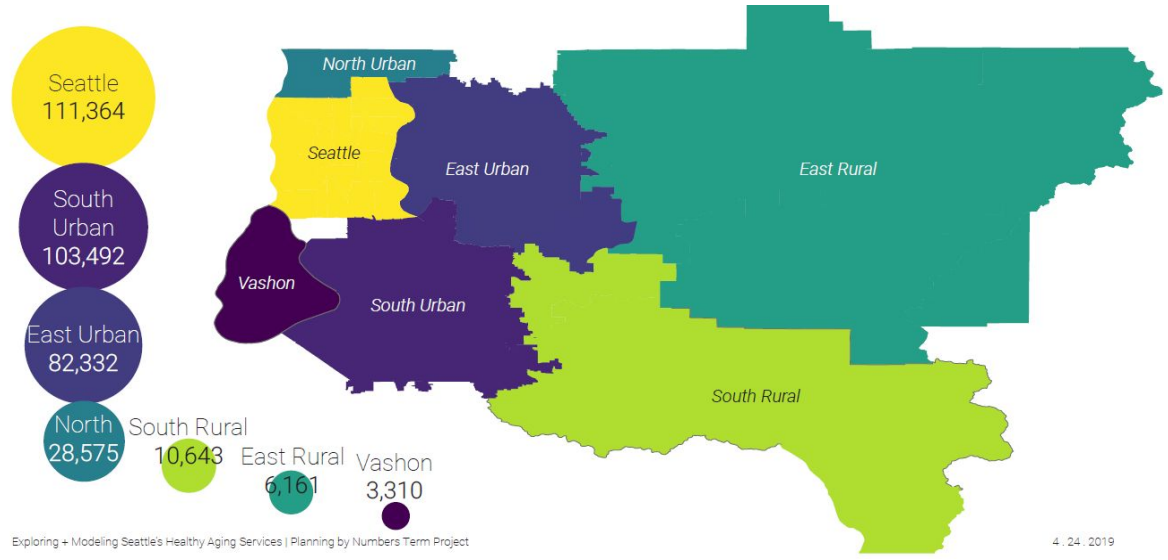



***Figure 10*** *- ADS Area of Analysis by Seattle's King County Area/Neighborhood Use Frequency Aggregated by Sub-regional Distribution. The bubbles on the side are combined counts of use.*
*Note: the following neighborhoods fall in the yellow "Seattle" catchment area - SW Seattle, SE Seattle, Queen Anne, North Seattle, NE Seattle, Lake Union, Duwamish, Downtown, Delridge, Central Seattle, Capitol Hill, and Ballard.*




***Client Demographics***

It was important to conduct a demographic examination of the ADS client base to craft a holistic understanding of who, of the aging population, was requesting these services and supports. Per the visualizations, it is apparent that the majority of reporting ADS clients fall in the age range of 65-79, with the individual age bracket 70-74 possessing the singular highest age capture; see ***Figure 11***. Additionally, approximately 61% of users were described as racially white. The second largest racial group, black, constituted about 22% of the clients' racial makeup; see ***Figure 12***. Finally, the team felt that since ADS support services are subsidized through a mixture of federal and state funding mechanisms, client income

might be skewed toward the lower income ranges as many of the clients are enrolled in Medicare and Medicaid services. Per the data analysis and visualizations, this assumption was confirmed. Approximately 72% (9,990 people) fell into the very low-income range; see ***Figure 13***.[10]

The ADS client data demographics differ slightly from the metropolitan region of Seattle. In Seattle, the population is 68.6% white, 14.5% asian, 7.1% black, 9.8% other; see Appendix ***Figure A***. A larger proportion of the overall black population, and less of the overall asian population generally represented in the region are utilizing the specific ADS services. Finally, some elderly minority groups, immigrants, or refugees may not be as represented in the ADS reported client base because their limited english speaking abilities could diminish service use rates. This particular group subset may be unaware that these services are available to them, or are uncomfortable accepting the assistance. Additionally, healthy distrust of government-funded services, and personal information privacy may offer another logic as to why these groups are not as well represented in the data.[11]

***Figure 11*** - *Age Range Grouping of Clients*

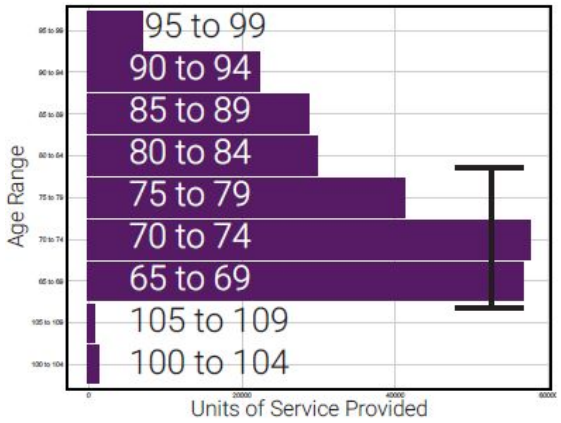



***Figure 12*** - *Racial Grouping of Clients*

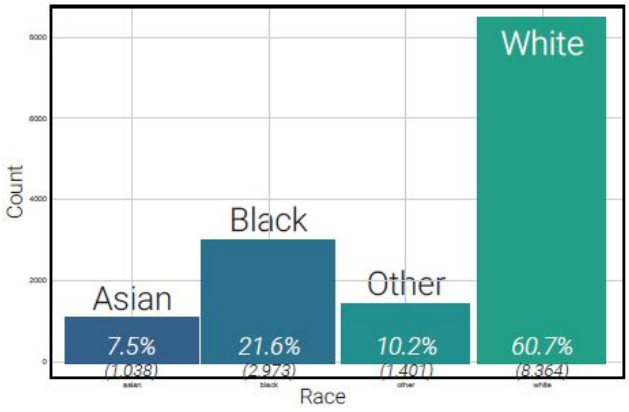
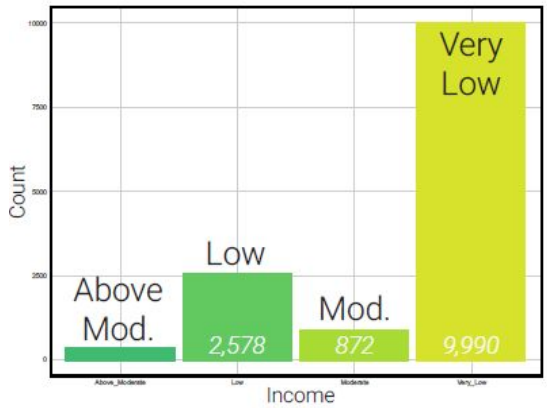



***Figure 13*** - *Income Grouping of Clients*




[10] "Very Low Income" per the Seattle ADS metadata captures any client that reports an income of less than 30% below the area's median income.
[11] Colin Macarther, "How People Learn to Navigate Government Services," Government, DigitalGov, 43:04 - -0400 400AD, /2016/03/03/how-people-learn-to-navigate-government-services/.

***Variable Overlap to Note***

The team explored the service user descriptor and Activities of Daily Living (ADLs) help binary variables; see ***Appendix Figure B*** for a full binary variable list. In reviewing the dummy sequences, certain relationships were clearly visible. For example, within the descriptor binaries if the client was disabled they were also more likely to live alone, be a veteran, and have a nutritional risk tendency. Additionally, in observing the ADL help needed, many clients did not require help with money but did desire help shopping.

***Selecting Variables***

Development of four regression models enabled the used of the predictive power of variable selection. This fundamental predictive power allowed the team to efficiently identify and select the variables that would lend themselves to generating an optimum regression.

The team used fact-based theories and grounded data assumptions to inform coefficient incorporation. Some variables that were included based on this intuition, for example, were the clear link between "help driving" and "help getting to places" and "help transferring". Each of these service "help" examples require movement assistance, and are ostensibly connected. Finally, since the team's overarching goal was to develop four predictive models, the significance of the variables within the regression (p-values) enable the variable inclusion.

**Regression Outcomes + Initial Foray Into Predictive Modeling**

***Overview + Key Points***

Four separate regressions were built to evaluate the relevance of the binary variables in predicting the known outcomes of service use: Case Management, Nutrition, Transportation, and Information & Assistance. There were 13,776 total observations in the 2015 cleaned dataset. Since Case Management (3,716 cleaned observations) and Nutrition (8,781 cleaned observations) had the largest number of individual client observations, and are therefore less rare, the regression results appear more relevant.[12] Additionally, it was assumed that the next steps for predictive modeling will enable more accurate model forecasting. This will be explored in the Predictive Modeling (below) section of the report.

In developing each of the four binomial logit models the team used a thoughtful "kitchen sink" approach - throwing "everything but the kitchen sink" into the regression to craft an unrefined snapshot of statistical patterns. All independent variables identified by the team as plausibly meaningful were incorporated into each of the first model attempts. After each model was constructed, the results were summarized to garner a sense of the deviance residual distribution, and the displayed p-value significance of each coefficient. Variables with insignificant p-values were removed from the second-round model evolution, and the models were re-run and re-summarized. Each of the four binomial logit regressions are described below.

[12] Case Management (108,827 initial observations), Nutrition (78,102 initial observations), Information and Assistance (25,923 initial observations), and Transportation (9,960 initial observations)

***Case Management Regression, Density Plot + ROC Curve*** [13]

A stargazer binomial regression table was developed via the use, and then refinement of a "kitchen sink" regression, and then the refinement of coefficients into a more meaningful lean regression model. ***Figure 14*** below highlights the regression summary table for the final Case Management prediction model developed for this report. Each coefficient has an associated p-value (indicator of variable significance), as well as an odds ratio value (illustrated by the numerical value within the parentheses). Each of the coefficients exhibit strong significance to the Case Management logistic regression, as designated by the presence of "**" or "***" adjoining individual variable outputs.[14]

In order to further expand on the coefficient interpretation within the model, the team investigated the reported odds ratio output listed in the summary table; see ***Figure 14***.[15] [16] The subsequent bullet points consider the respective variable's implications and their relationship to Case Management use.

- *Speaks Limited English (speak_english)*: this coefficient is a binary variable that assigns a value of 1 for those individuals who speak limited English. It is associated with an odds ratio of 0.091, this implies that a user who speaks limited english has a 9.1% greater chance of using the case management service than a user who speaks english well.
- *Is a Veteran (is_veteran)*: similar to above, this coefficient is also a binary variable. Unlike the variable above, since this coefficient displays a negative output, the odds ratio is interpreted as - a person who is a veteran as a 6.9% decreased likelihood of using the case management service.

The two odds ratio interpretations in conjunction with understanding the p-value significance factors highlight how the team went about interpreting the regression summary tables for each of the four models.

***Figure 14*** - *Case Management Binomial Logistic Regression Model Summary Table*

| | Dependent variable: |
|---|---|
| | Used Case Management |
| AgeRange105 to 109 | -0.559 (0.468) |
| AgeRange65 to 69 | -0.809** (0.337) |
| AgeRange70 to 74 | -0.980*** (0.337) |
| AgeRange75 to 79 | -0.936*** (0.339) |
| AgeRange80 to 84 | -0.682** (0.338) |
| AgeRange85 to 89 | -0.750** (0.339) |
| AgeRange90 to 94 | -1.095*** (0.342) |
| AgeRange95 to 99 | -0.801** (0.371) |
| dummy_nutriRisk | 0.585*** (0.067) |
| speak_english | 0.385*** (0.091) |
| has_children | 0.563*** (0.096) |

[13] Note that the Case Management regression section contains the most detailed description of the processes, analyses, and conclusions the team completed for each of the four models. The subsequent models underwent similar scrutiny. Due to the length of the paper we thought it pertinent to only thoroughly highlight in one of the regression sections, not all.

[14] "*" = p-value less than 0.1 (<0.1); "**" = p-value less than 0.05 (<0.05); "***" = p-value less than 0.01 (<0.01)

[15] Odds ratios are denoted in the regression summary table via the numerical value listed within the parentheses "( )"

[16] Note that these regression summaries were developed for all four of the most significant models, so the p-value and odds ratio implications are note as explicitly stated in the subsequent sections. All summary factors maintain the same summary magnitude of importance throughout this paper unless otherwise stated.

| | | |
|---|---|---|
| is_disabled | 1.418*** | (0.071) |
| is_veteran | -0.344*** | (0.069) |
| help_toileting | 0.222*** | (0.071) |
| help_gettingplaces | 0.510*** | (0.062) |
| help_transfering | 0.697*** | (0.082) |
| help_dressing | 0.503*** | (0.071) |
| help_medical | 0.585*** | (0.083) |
| help_cooking | -0.730*** | (0.115) |
| help_chores | 0.547*** | (0.123) |
| help_phoning | -0.299*** | (0.069) |
| Constant | -1.911*** | (0.333) |
| Observations | | 13,776 |
| Log Likelihood | | -5,779.858 |
| Akaike Inf. Crit. | | 11,603.720 |

Note: *p<0.1; **p<0.05; ***p<0.01

Once the team determined the regression had reached its maximal reliability, the dataset was divided into a train and a test set. Respectively 70% and 30% of the total observations. The model was trained using the observations in the train set, and the tests were conducted on the testset observations. This method was used in all of the subsequent models.

Density plots were developed to understand the power of predicting Case Management's use of service probability for the testset; see ***Figure 15***. The outcomes for observations where the user used the service are plotted in purple and the probabilities for users that did not use Case Management were plotted in yellow. These density plots visualize the predicted probability distribution of data in the 2015 one-year time period. Observing the peaks of the density plot help identify predicted value concentrations within the interval.

Once an understanding of the density plot was developed, the team built a Receiver Operating Characteristic curve (ROC curve).[17] ROC curves visually plot the true positive rate (sensitivity) of occurrence against the false positive occurrence rate (specificity). In reviewing the visual output, it is important to note that the closer the ROC curve is to the upper left corner of the plot, the higher the overall accuracy of the model is perceived to be. The Case Management ROC curve output is reviewed below to determine the accuracy of the model; see ***Figure 16***. The output displays that the Area Under the Curve (AUC) for the best Case Management model developed is .8447 (84.47%). AUC measures how well the parameter can distinguish between two outcomes. An AUC of 84.47% represents that this model possesses reasonably strong predictive power and accuracy. The ROC curve slope highlights the total accurate Case Management use predictions possible based on the adjusted thresholds.

The next step was to define a threshold in the predictions to classify them. Since the absolute value of the predictions will depend on how rare or how often an event is within a dataset, each model should be individually calibrated. By analyzing the density plot of this model, it is visible that the sweet spot - the

[17] Density plots and ROC curves were developed for each of the four regression models discussed in further detail within each respective section..

point at which the model delivers a good balance in predicting both true and false occurrences - is close to 25%. After a series of optimization tests, a threshold of 35% was established to build a confusion matrix. This means that every observation with a prediction over 35% was classified as an individual who will most likely use the service. The classified results are then compared to the observed results in the confusion matrix. At the 35% threshold, the Case Management's model accuracy rating was reported at 80%, sensitivity at 75%, and specificity at 81%.[18] An accuracy rating greater than or equal to 80% signifies a relatively strong model, so the Case Management model is deemed statistically robust.

***Figure 15*** *- Density Plot of Test Set Predicted Probabilities for Did Not Use (yellow) vs Used (purple) Case Management*

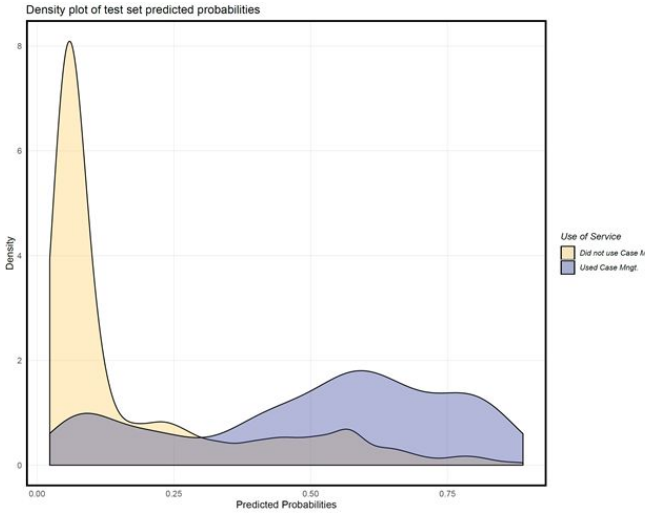
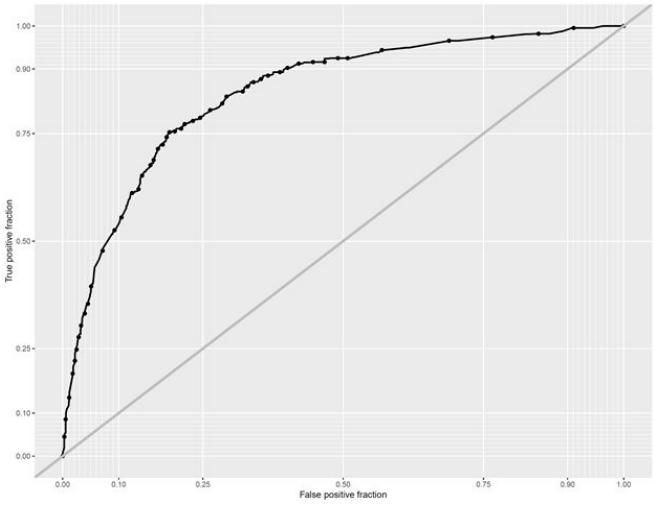



***Figure 16*** *- ROC Curve for Case Management; False Positive and True Positive Likelihood Distribution*




***Nutrition Regression, Density Plot + ROC Curve***

The Nutrition regression was a particularly interesting model since Nutrition is the most used service of all four analyzed service categories. Nearly 64% of all clients within the dataset recorded having used the Nutrition service offering. Per the tremendous usage rate, the team has assumed that this heavy utilization resulted in overall higher probabilities illustrated within the Nutrition model. Unlike the regression models associated with Case Management (detailed above) or Transportation (detailed later in the report), the Nutrition model, seen in ***Figure 18***, produces very high predictions amid people who reported using the Nutrition service. Conversely, clients who did not use the Nutrition service displayed lower predicted probabilities, but a still relatively moderate chance of potential future use.

The Nutrition model ROC curve AUC presented as .80. Similar to the Case Management model above, the Nutrition model possesses prudently strong predictive power and accuracy; see ***Figure 19***. Additionally, the Nutrition model's confusion matrix threshold was set at 70%. At this threshold the confusion matrix detailed that the model output assumed 72% accuracy, 70% sensitivity, and 77% specificity.

[18] Confusion matrix outputs highlight the explicit accuracy of the best fit Case Management model. A confusion matrix contrasts predicted results (in this case for 2016) against those those in the cleaned dataset (in this case 2015).

***Figure 17*** *- Nutrition Binomial Logistic Regression Model Summary Table*

| | Dependent variable: |
|---|---|
| | Used Nutrition |
| IncomeLow | -1.312*** (0.270) |
| IncomeModerate | -1.157*** (0.283) |
| IncomeVery_Low | -1.405*** (0.267) |
| dummy_livealone | -0.216*** (0.051) |
| dummy_nutriRisk | -0.631*** (0.072) |
| speak_english | -0.271*** (0.092) |
| has_children | -0.305*** (0.093) |
| is_homeless | -1.806*** (0.635) |
| is_disabled | -1.314*** (0.062) |
| live_outerbounds | -0.865*** (0.164) |
| is_veteran | 0.313*** (0.065) |
| help_toileting | -0.124* (0.074) |
| help_walking | 0.303*** (0.061) |
| help_gettingplaces | -0.458*** (0.063) |
| help_transfering | -0.620*** (0.087) |
| help_dressing | -0.307*** (0.072) |
| help_bathing | -0.114* (0.068) |
| help_medical | -0.567*** (0.089) |
| help_cooking | 0.683*** (0.103) |
| help_shopping | -0.167* (0.090) |
| help_driving | -0.255** (0.103) |
| help_phoning | 0.249*** (0.071) |
| Constant | 3.206*** (0.280) |
| Observations | 12,432 |
| Log Likelihood | -6,328.577 |
| Akaike Inf. Crit. | 12,703.160 |

Note: *p<0.1; **p<0.05; ***p<0.01

***Figure 18*** *- Density Plot of Test Set Predicted Probabilities for Did Not Use (yellow) vs Used (purple) Nutrition*

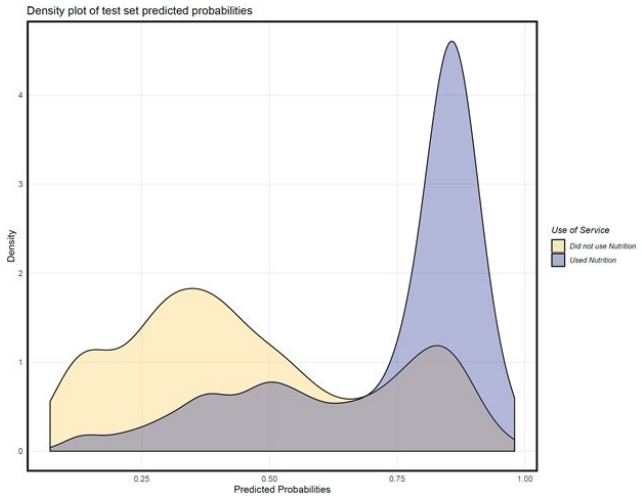
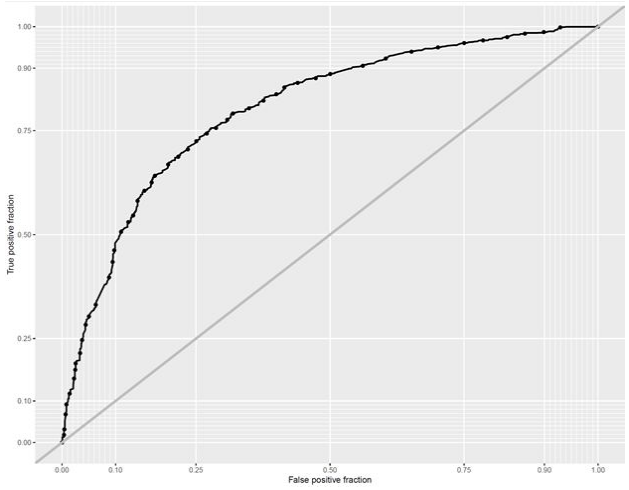



***Figure 19*** *- ROC Curve for Nutrition; False Positive and True Positive Likelihood Distribution*




***Transportation Regression***

In this dataset, Transportation is a more rare event. The density plot shows that the predictions are overall low. The highest likelihood of using Transportation in this dataset is around 40%. The model performs well for correctly assigning low likelihood for those who did not use transportation. Unfortunately, it gives also low likelihood for many individuals who used it; see ***Figure 21*** .

The Transportation confusion matrix was built at the low threshold of probability classification of 8% (0.08). This threshold level yielded an accuracy of 87%, sensitivity of 56%, and specificity of 89%. Most likely due to fewer users electing to use the Transportation service, and therefore a lower sample size than the numbers associated with Case Management and Nutrition, the Transportation model did not perform quite as well. Particularly, it proved difficult to procure a high sensitivity yield from this model. The low associated sensitivity is likely due to the large overlap in low predictions among clients who used and did not use Transportation (0-10% likelihood).

Although this model demonstrates some fundamental flaws, the ROC curve produced is relatively impressive. The listed AUC is .81; higher than that of Nutrition and Case Management; see ***Figure 22***.

***Figure 20*** *- Transportation Binomial Logistic Regression Model Summary Table*

| | Dependent variable: | |
|---|---|---|
| | Used Transportation | |
| GeographicLocation East Urban | 0.710 | (0.544) |
| GeographicLocation North Urban | 0.808 | (0.541) |
| GeographicLocation Ballard | 0.983* | (0.564) |
| GeographicLocation Capitol Hill | -0.032 | (0.659) |
| GeographicLocation Central Seattle | -1.891** | (0.895) |
| GeographicLocation Delridge | -1.187 | (0.734) |
| GeographicLocation Duwamish | 0.770 | (0.610) |
| GeographicLocation Lake Union | 1.483** | (0.591) |
| GeographicLocation NE Seattle | 0.213 | (0.571) |
| GeographicLocation North Seattle | -1.419* | (0.793) |
| GeographicLocation Queen Anne | 0.432 | (0.578) |
| GeographicLocation SE Seattle | -1.267** | (0.606) |
| GeographicLocation SW Seattle | 0.763 | (0.573) |
| GeographicLocation South Rural | -13.154 | (200.680) |
| GeographicLocation South Urban | 0.208 | (0.535) |
| GeographicLocation Vashon | -13.649 | (532.797) |
| dummy_livealone | 1.077*** | (0.124) |
| is_disabled | 1.385*** | (0.111) |
| live_outerbounds | 1.904*** | (0.223) |
| help_walking | -0.699*** | (0.109) |
| help_medical | -1.117*** | (0.179) |
| help_shopping | 0.686*** | (0.189) |
| help_chores | -0.993*** | (0.263) |
| help_driving | 0.562** | (0.261) |
| raceblack | 0.533* | (0.293) |
| raceother | -0.261 | (0.343) |
| racewhite | 0.786*** | (0.264) |
| Constant | -5.227*** | (0.627) |
| Observations | 12,432 | |
| Log Likelihood | -1,870.876 | |
| Akaike Inf. Crit. | 3,797.752 | |
| Note: | *p<0.1; **p<0.05; ***p<0.01 | |

***Figure 21*** *- Density Plot of Test Set Predicted Probabilities for Did Not Use (yellow) vs Used (purple) Transportation*

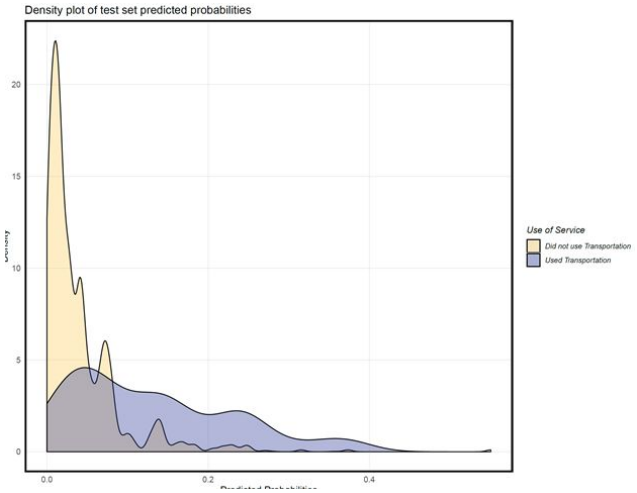
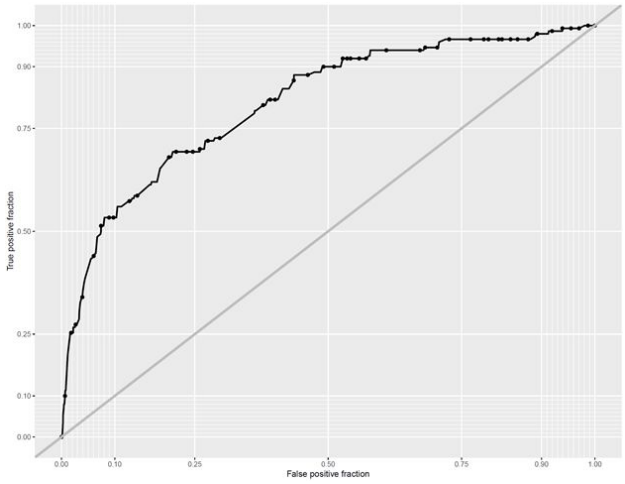



***Figure 22*** *- ROC Curve for Transportation; False Positive and True Positive Likelihood Distribution*




***Information and Assistance Regression***

The Information and Assistance Regression proved the most difficult model to predict. As the density plot in ***Figure 24*** shows, there is a large overlap of probabilities for clients who used the Information and Assistance service, and those clients that did not. This overlap is visualized by the blended yellow and purple density curves. Additionally, all associated probabilities are very low, spanning 20% at most.

By setting the confusion matrix threshold to 8%, the Information and Assistance model outputs an 80% accuracy reading, 38% sensitivity, and 83% specificity. Observing the ROC curve in ***Figure 25,*** the plotted arch is very flat and tracks more closely with the 50/50 boundary indicating that the Information and Assistance regression is mildly more predictive than simply opting to flip a coin (e.g., 50/50 chance). Additionally, the AUC is reported at .67 which further deduces that this model maintains lower predictive power and accuracy.

***Figure 23*** *- Information and Assistance Binomial Logistic Regression Model Summary Table*

| | Dependent variable: |
|---|---|
| | Used Information and Assistance |
| DISTANCE_MILES | -0.076*** (0.010) |
| dummy_hispanic | 0.298* (0.168) |
| IncomeLow | 1.790*** (0.510) |
| IncomeModerate | 1.267** (0.532) |
| IncomeVery_Low | 1.240** (0.508) |
| speak_english | 0.621*** (0.136) |
| is_homeless | 2.159*** (0.484) |

| | | |
|---|---|---|
| is_disabled | 0.469*** | (0.097) |
| help_eating | -0.365*** | (0.133) |
| help_toileting | -0.223* | (0.126) |
| help_gettingplaces | -0.325*** | (0.101) |
| help_bathing | 0.248** | (0.112) |
| help_shopping | -0.404*** | (0.129) |
| Constant | -3.495*** | (0.518) |
| Observations | 12,432 | |
| Log Likelihood | -2,540.716 | |
| Akaike Inf. Crit. | 5,109.433 | |

Note: *p<0.1; **p<0.05; ***p<0.01

***Figure 24*** *- Density Plot of Test Set Predicted Probabilities for Did Not Use (yellow) vs Used (purple) Information and Assistance*

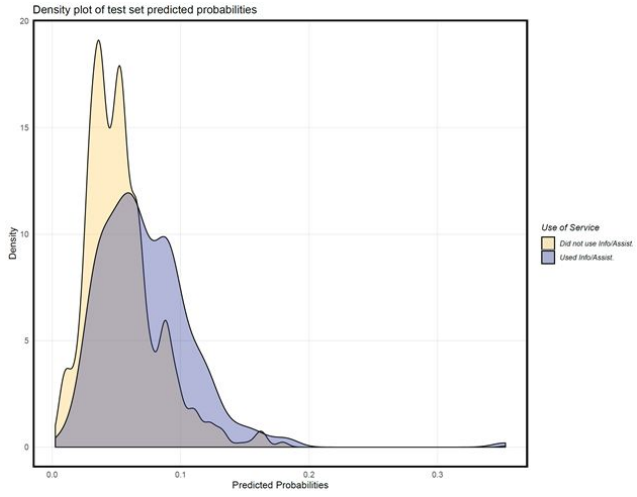
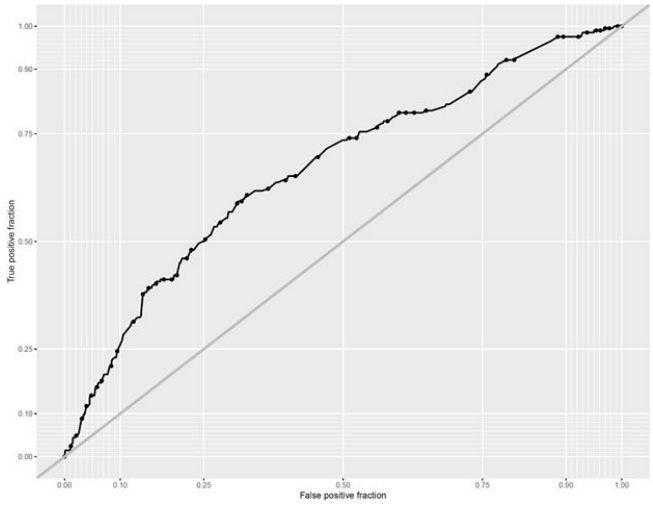



***Figure 25*** *- ROC Curve for Information and Assistance False Positive and True Positive Likelihood Distribution*




**Further Exploration of Predictive Model Outcomes for Our Best Fit Model: Case Management**

***Case Management Predictive Model Investigation***

After the final regression for each of the four models the team selected the best model (Case Management) to further investigate its predictive power. The first step was to developed an association matrix analysis to understand the relationships amongst binary variables, and how these relationships might impact standard error.[19] The association table output highlighted the multicollinearity between coefficients. Additionally, although the Association Table, displayed in ***Figure 26,*** shows high association between independent variables at some intersections, when the team attempted to remove

[19] The team attempted using the association table sine the VIF doesn't catch association between dummy variables

these variables from the Case Management models, the overall model output, and specifically the ROC curve did not improve. Therefore, the initial variables used were kept. In addition to the strong binary-to-binary affiliation, the association matrix's numerous low p-values highlight that our dependent variable of Case Management, and the independent variables are highly associated. It is important to note that allowing these collinear predictors to remain has the potential to skew the model results, but the team felt that keeping the specific indicators used with in the Case Management model was crucial based on the theoretical client service needs. The team selected to use the large model and watch for heighted skewing, as opposed to the reduced model, due to the hypothesized importance of the variables.

All p-values listed in the association table displayed significance except the following combinations: *disabled x has children*, *help getting places x speaks English*, *help transferring x speaks English*, *help cooking x speaks English*, *help with chores x speaks English*. These lower reported association values might first be due to the general disabled population typically having fewer children than those people who are not disabled; therefore the low presence of association is logical.[20] The second grouping of reduced association is related to the coefficient of *speaks English*. The team assumes that *speaks English* might have a suppressed association based on the lower number of reported clients who have limited English speak capabilities highlighting that this group is not as well represented within the dataset. Another theory is that maybe little verbal communication is actually required to access these services linked with *speaks English*, and so there is a lower association output.

***Figure 26*** - *P-Value Association Table*

| P-VALUES | used_case_mngm | AgeRange | dummy_nutriRisk | speak_english | has_children | is_disabled | is_veteran | help_toileting | help_gettingplaces | help_transfering | help_dressing | help_medical | help_cooking | help_chores | help_phoning |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| used_case_mngm | | | | | | | | | | | | | | | |
| AgeRange | 0.0000 | | | | | | | | | | | | | | |
| dummy_nutriRisk | 0.0000 | 0.0000 | | | | | | | | | | | | | |
| speak_english | 0.0000 | 0.0000 | 0.0000 | | | | | | | | | | | | |
| has_children | 0.0000 | 0.0000 | 0.0000 | 0.0000 | | | | | | | | | | | |
| is_disabled | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.4521 | | | | | | | | | | |
| is_veteran | 0.0000 | 0.0000 | 0.0014 | 0.0000 | 0.0000 | 0.0000 | | | | | | | | | |
| help_toileting | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | | | | | | | | |
| help_gettingplaces | 0.0000 | 0.0000 | 0.0000 | 0.3215 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | | | | | | | |
| help_transfering | 0.0000 | 0.0000 | 0.0000 | 0.1064 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | | | | | | |
| help_dressing | 0.0000 | 0.0000 | 0.0000 | 0.0162 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | | | | | |
| help_medical | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | | | | |
| help_cooking | 0.0000 | 0.0000 | 0.0000 | 0.0643 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | | | |
| help_chores | 0.0000 | 0.0000 | 0.0000 | 0.2706 | 0.0001 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | | |
| help_phoning | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | |

[20] "Disability & Socioeconomic Status," Government, https://www.apa.org, May 2019, https://www.apa.org/pi/ses/resources/publications/disability.

In observing the chi-square matrix output; seen in ***Figure 27***, the matrix output detailed numerous high chi-squared values (>1000). An item of interest is that the binary variables that represent clients that require help with chores (*help_chores*) and clients who are disabled (*is_disabled*) appear to be associated with all other help variables. This chi-squared outcome makes sense since the likelihood of a client who needs help with one type of daily task would theoretically also be linked to other daily service needs, especially when considering the presence of a disability factor. The binary variables representing clients who have limited English speaking capabilities (*speak_english*), and those who have children (*has_children*) appear to be the coefficients that are the least associated with binaries. Similar to the coefficient removal test inspired by the P-Value Association Matrix, the team tested the elimination of variables displaying suppressed levels of importance within the chi-squared matrix. Per the ROC curve AUC output summary table, displayed in ***Figure 28***, the removal of this variables in question did not improve the Case Management model. Therefore, the team decided to keep the original Case Management coefficients and maintain the superior AUC.

***Figure 27*** - *Chi-squared Matrix*

| CHI^2 | used_case_mngm | AgeRange | dummy_nutriRisk | speak_english | has_children | is_disabled | is_veteran | help_toileting | help_gettingplaces | help_transfering | help_dressing | help_medical | help_cooking | help_chores | help_phoning |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| used_case_mngm | | | | | | | | | | | | | | | |
| AgeRange | 94 | | | | | | | | | | | | | | |
| dummy_nutriRisk | 1,207 | 64 | | | | | | | | | | | | | |
| speak_english | 65 | 116 | 265 | | | | | | | | | | | | |
| has_children | 87 | 158 | 25 | 140 | | | | | | | | | | | |
| is_disabled | 3,030 | 403 | 1,236 | 104 | 1 | | | | | | | | | | |
| is_veteran | 190 | 734 | 10 | 33 | 28 | 384 | | | | | | | | | |
| help_toileting | 1,691 | 104 | 548 | 18 | 110 | 2,563 | 81 | | | | | | | | |
| help_gettingplaces | 703 | 170 | 2,415 | 1 | 39 | 650 | 34 | 134 | | | | | | | |
| help_transfering | 3,367 | 208 | 1,673 | 3 | 52 | 6,191 | 207 | 5,099 | 694 | | | | | | |
| help_dressing | 2,022 | 92 | 570 | 6 | 187 | 3,085 | 178 | 7,192 | 259 | 5,250 | | | | | |
| help_medical | 3,106 | 189 | 1,456 | 29 | 19 | 6,097 | 158 | 4,240 | 778 | 9,021 | 9,021 | | | | |
| help_cooking | 2,235 | 268 | 1,967 | 3 | 23 | 5,238 | 125 | 3,120 | 3,628 | 6,833 | 6,833 | 6,839 | | | |
| help_chores | 2,644 | 287 | 2,155 | 1 | 16 | 6,197 | 151 | 3,126 | 3,583 | 7,255 | 7,255 | 6,997 | 11,701 | | |
| help_phoning | 1,401 | 146 | 484 | 191 | 113 | 2,747 | 75 | 4,273 | 464 | 4,613 | 4,613 | 6,155 | 3,781 | 3,773 | |

***Figure 28*** - *ROC Curve AUC Summary Table*

| | |
|---|---|
| Original ROC: | 0.8447 |
| | |
| Took Out: | ROC Curve: |
| | |
| help_chores | 0.8444 |
| is_disabled | 0.8255 |
| help_chores + is_disabled | 0.8238 |
| speak_english | 0.8442 |
| has_children | 0.8437 |
| All above | 0.8194 |
| AgeRange | 0.8421 |

**Removing vars does not improve the model**

***K-Fold Cross-Validation***

To estimate the skill of the model on a new data set, the team decided to perform a 100-fold cross validation procedure. This cross-validation enables a resampling of data to evaluate machine learning modelling capabilities and accuracy with a limited data sample.[21] Per the cross-validation, the overall model output accuracy was 80%. The histogram displayed in ***Figure 29***, highlights that the accuracy of the k-fold test has a normal distribution and a particularly large range (approximately 74% to 88%).

The team thought it important to test the best performing model, the 2015 Case Management model, on 2016 data to begin to develop a comprehension for how these predictions inform real ADS service outcomes. ***Figure 30*** highlights the test density plot of the model on 2016 data. Overall, the model performed well. Assessing the confusion matrix output, the team used the same threshold previously honed for the test set (35%), this threshold enabled an 82% accuracy rating, 76% sensitivity, and 84% specificity; see results in ***Figure 31.***

In reviewing the output results via a more theoretical basis, the team recognized that the predictive power of the 2015 dataset-trained model might retain some paradoxical outcomes. Aging users who relied on ADS services in 2015 were highly likely to also utilize the same (or more) services in 2016. This troublesome inference spurred the team to construct and run some exploratory simulations.

***Figure 29*** *- Histogram of 100-Fold Cross-Validation Accuracy Output*

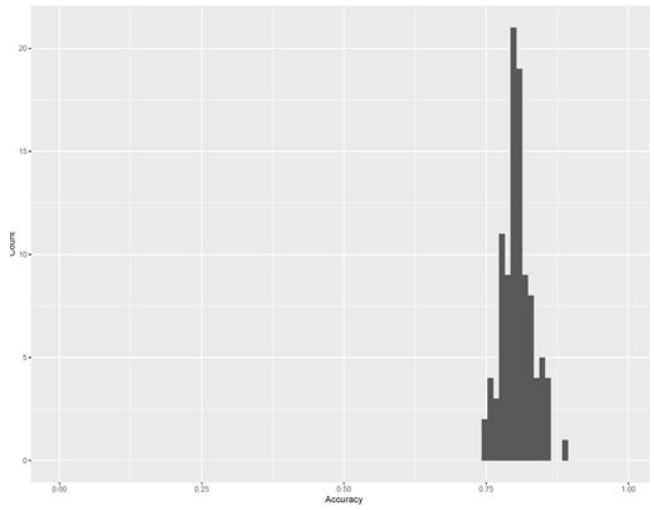
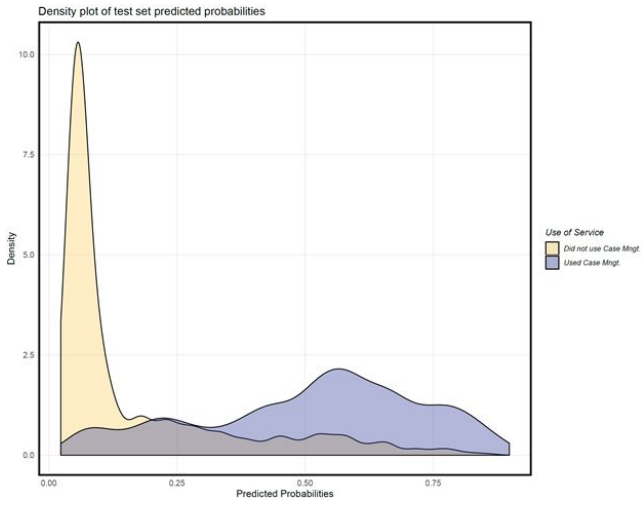
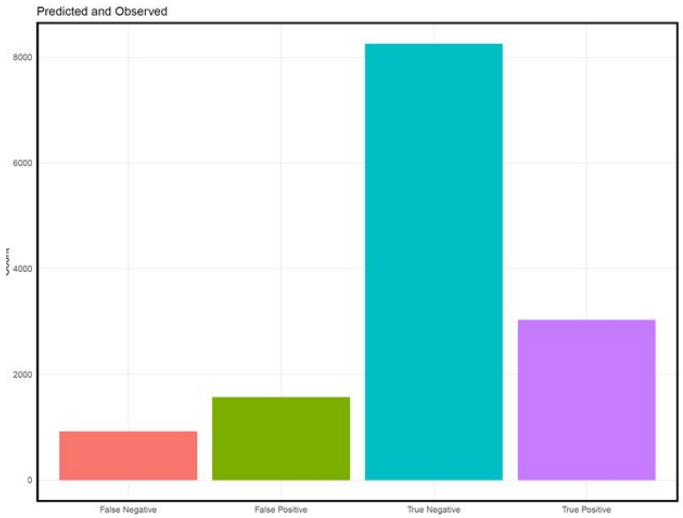



***Figure 30*** *- Case Management Testing of Model's Predictive Capabilities (tested on 2016 data)*




[21] The common k-fold (in this case 100-fold) cross-validation is: (1) rearrange the dataset randomly; (2) divide the data into 100 groups; (3) for all unique groups: employ a specific test data set, use surplus data to "train" the set, fit the model to the training set and assess the test set, preserve the evaluation score and scrap the model. (4) Finally, summarize the skill of the model using the sample of model evaluation scores.

***Figure 31*** *- Histogram of Confusion Matrix True -/+ & False -/+*




***Simulation Testing***

The first simulation run established a fabricated 50% increase in the Case Management disabled population. To create this speculative increase 76% of the observations were randomly allocated as "disabled".[22] As seen in ***Figure 32***, the density plot highlight a significant portion of the population did not use Case Management, and received suppressed probabilities. The original Case Management model suffered a significant increase in probability of using the Case Management service. This is easily observed in ***Figure 32*** by the second yellow peak of the plot around the 20% probability threshold. Within this simulation, the average probability of using Case Management increased to 30% (27% in the original model), and the median increased to 23% (12% originally).

The second simulation, seen in in ***Figure 33***, presumed a five-times larger population that had limited English-speaking capabilities (e.g., likely to be immigrants, refugees, or elderly visiting family members). The Case Management density plot appears to be unaffected by this alteration; its appearance aligns very closely with the original output. The average probability of utilizing Case Management increased by only 1% relative to the original model (from 27% originally to 28% with the respective increases), and the median increased by only 2% (from 12% to 14%). This model serves as proof of concept that having enough information regarding the senior population can lead to better understanding of the demand for these healthcare and wellbeing oriented services.

[22] This simulation was run on the 2015 dataset.

***Figure 32*** *- Density Plot of 2015 Case Management Predicted Probabilities with Disabled Simulation Inclusion*

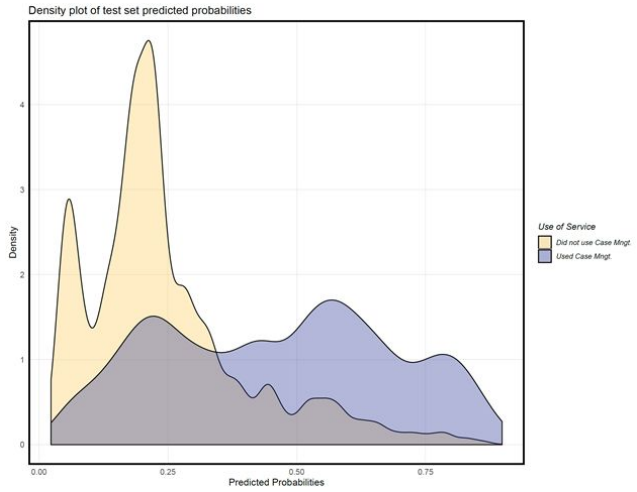
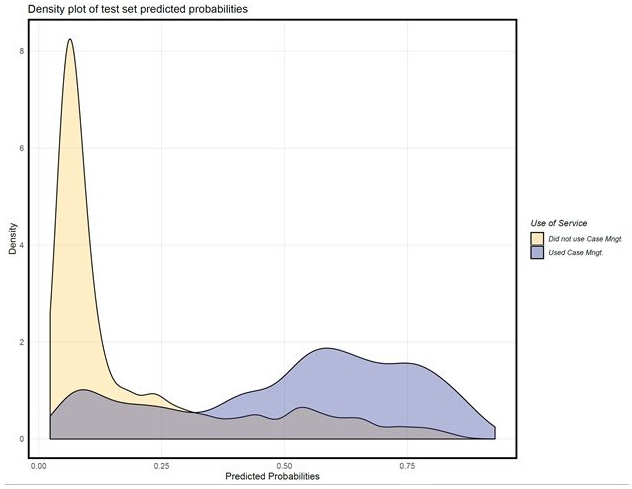



***Figure 33*** *- Density Plot of 2015 Case Management Predicted Probabilities with Limited English Simulation Inclusion*




***Conclusion***

The predictive powers of the four models -- Case Management, Nutrition, Transportation, and Information and Assistance -- have strong real world implications. Each of the models performed slightly differently from one another, with Case Management lending itself as the strongest predictive model, with the best ROC curve (~.85 area under the curve), out of the bunch. With these high accuracy outputs, the local government of King County Seattle could rely on these models to forecast increasing budgetary demands in the region. Fiscal changes are expected to rise as the baby boomer age wave occurs. This age wave will place additional strain on the ADS system, so rigorous futurecasting is critical. Secondly, this model could be used to identify particular Seattle neighborhoods with low-income, non-english speaking members, and specific outreach efforts could target each community's unique demographic service needs (e.g., translators could be employed, tailored culturally-sensitive aging services, the addition of service offerings). By identifying underserved, vulnerable communities, this model lends itself to behave as an equity tool for the aging population. Ensuring each community's elders are cared for in a dignified manner later inlife. Finally, since these factors were debunked as being highly informed by definitive client specific needs, and not spatially determined, this model could conceivably evolve to be used and tested in a different place. Per size, demographics, landscape, and population, potential cities that could use and expand upon this model are Portland, Oregon; San Francisco, CA; Minneapolis, Minnesota; and Madison, Wisconsin.[23] [24]

To further improve upon this model, the team would encouraging working with the ADS organization to craft an fact-based understanding of the survey, data collection, and specific reporting process. This additional information on how the binary variables are collected and exactly defining what they represent

[23] "R/SeattleWA - Which Larger Cities (Pop. 1M+) Are Most Similar to Seattle?," Posts, reddit, May 2019, https://www.reddit.com/r/SeattleWA/comments/6hwxcy/which_larger_cities_pop_1m_are_most_similar_to/.
[24] A similar amount of information pertaining to the elderly population would be required to run a similar model

(e.g., the explicit difference between “transferring” versus “getting to places”) will empower a superior model. Overall, this report acts as an initial exploration toward developing a more comprehensive understanding of Seattle’s aging population’s current and future service needs. The models’ most likely real world applications are realized as a budgetary forecasting mechanism, elderly equity tool, and model informant for like-sized and data-rich heavy counterpart cities.

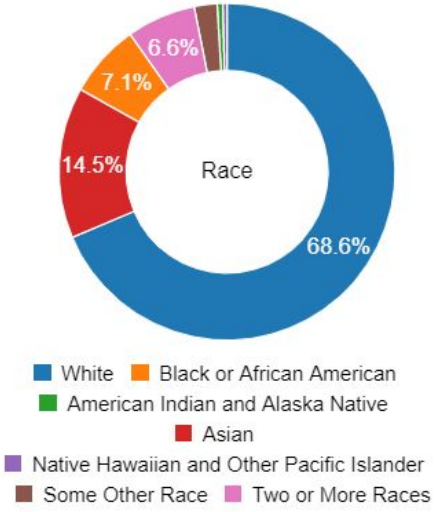
## Appendix

***Appendix Figure A*** **-** Seattle Metropolitan Region by Race (2017 ACS data)

**Population by Race**

| Total | Hispanic | Non-Hispanic |
|---|---|---|

| Race | Population ▾ |
|---|---|
| White | 472,347 |
| Asian | 99,728 |
| Black or African American | 48,884 |
| Two or More Races | 45,657 |
| Some Other Race | 15,155 |
| American Indian and Alaska Native | 3,799 |
| Native Hawaiian and Other Pacific Islander | 2,675 |



***Appendix Figure B -*** ADS dataset categorical variables (e.g., income, race, region) and binary variables (e.g., help, disabled, veteran, etc.)

| FIELD | *TYPE* & CODES | DESCRIPTION/QUESTION |
|---|---|---|
| ActivityID | *integer* | *Unique identifier for record in database* |
| ClientID | *integer* | *Unique identifier for client, internal to database* |
| GeographicLocation | *controlled text entry* | *Describes the geographic location of the client, based on provided zip code & City-defined areas. Zip codes included in areas assigned below.* |
| | East Rural | 98014, 98019, 98024, 98045, 98050, 98065, 98068, 98224, 98251, 98288 |
| | East Urban | 98004, 98005, 98006, 98007, 98008, 98009, 98015, 98027, 98029, 98033, 98034, 98039, 98040, 98052, 98053, 98059, 98073, 98074, 98075, 98083 |
| | North Urban | 98011, 98028, 98041, 98072, 98082, 98155, 98160 |
| | Seattle Neighborhoods: Ballard | 98107, 98117 |
| | Seattle Neighborhoods: Capitol Hill | 98102, 98112 |
| | Seattle Neighborhoods: Central Seattle | 98122 |
| | Seattle Neighborhoods: Delridge | 98106, 98126 |
| | Seattle Neighborhoods: Downtown | 98101, 98104, 98111, 98114, 98121, 98129, 98154, 98161, 98164, 98174, 98181, 98184, 98191 |
| | Seattle Neighborhoods: Duwamish | 98108, 98124, 98134 |
| | Seattle Neighborhoods: Lake Union | 98103 |
| | Seattle Neighborhoods: NE Seattle | 98125, 98133 |
| | Seattle Neighborhoods: North Seattle | 98105, 98115, 98145, 98177, 98185, 98195 |
| | Seattle Neighborhoods: Queen Anne | 98109, 98119, 98199 |
| | Seattle Neighborhoods: SE Seattle | 98118, 98144 |
| | Seattle Neighborhoods: SW Seattle | 98116, 98136, 98146 |
| | South Rural | 98010, 98022, 98025, 98038, 98051 |
| | South Urban | 98001, 98002, 98003, 98023, 98030, 98031, 98032, 98035, 98042, 98047, 98054, 98055, 98056, 98057, 98058, 98062, 98063, 98064, 98071, 98092, 98093, 98131, 98132, 98138, 98148, 98151, 98158, 98166, 98168, 98170, 98171, 98178, 98188, 98190, 98198, 98354 |
| | Vashon | 98013, 98070 |
| AgeRange | text | *5-year age ranges starting at "0 to 4." Age is based on current age at the time of the data pull.* |
| EthnicityCode, Ethnicity | text | *What is the client's ethnicity?* |
| | u | Unknown |
| | y | Hispanic or Lation |
| | n | Not Hispanic or Latino |
| RaceCode, Race | *integer* | *What is the client's race?* |
| | 0 | Unknown |

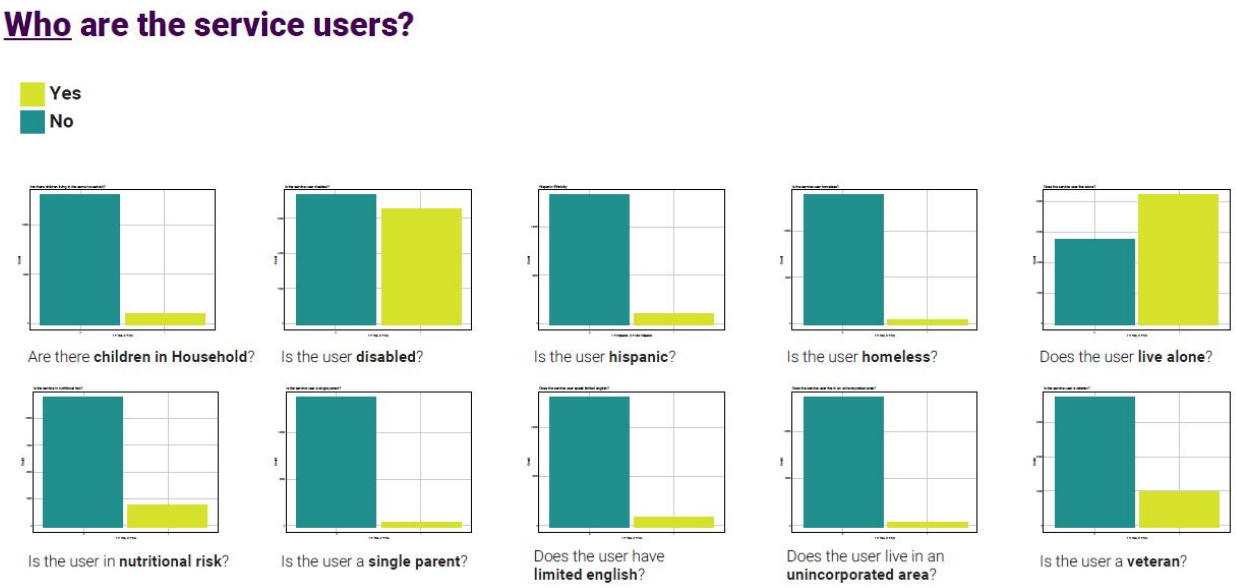
| | | |
|---|---|---|
| | 1 | American Indian or Alaska Native |
| | 2 | Asian, Asian-American |
| | 3 | Black, African-American, Other African |
| | 4 | hawaiian Native or Pacitic Islander |
| | 5 | Hispanic, Latino |
| | 6 | White or Caucasian |
| | 7 | Other |
| | 8 | Multi-Racial |
| IncomeCode, Income | *integer* | *Categorization of income, based on City Income Guidelines* |
| | 0 | Unknown |
| | 1 | Very Low (<30% Median) |
| | 2 | Low (<50% Median) |
| | 3 | Moderate (<80% Median) |
| | 4 | Above Moderate (>80% Median) |
| LiveAlone | *text* | *Does a client live alone?* |
| | u | Unkown |
| | y | Yes |
| | n | No |
| LimitedEnglish | *text* | *Does the client have limited proficiency in English?* |
| | u | Unkown |
| | y | Yes |
| | n | No |
| Language | *text* | *The primary langauge of the client* |
| NutritionalRisk | *text* | *Is a client at risk of poor nutritional health?* |
| | u | unknown |
| | n | No (nutrition assessment score below 6) |
| | y | Yes (nutritional assessment score 6 or greater) |
| | 0 | Good |
| | 1 | |
| | 2 | |
| | 3 | Moderate Nutritional risk |
| | 4 | |
| | 5 | |
| | 6 | High nutritional risk |
| | 7 | |
| | 8 | |
| | 9 | |
| SingleParent | *text* | *Is the client a single parent?* |
| | u | Unkown |
| | y | Yes |
| | n | No |
| HouseholdWithChildren | *text* | *Does the client live in a househodl with children under age 18?* |

| | | |
|---|---|---|
| | u | Unkown |
| | y | Yes |
| | n | No |
| Homeless | *text* | *Is the client homeless?* |
| | u | Unkown |
| | y | Yes |
| | n | No |
| DisabilityStatus | *text* | *Does the client have a disability?* |
| | u | Unkown |
| | y | Yes |
| | n | No |
| Unincorporated | *text* | *is the client part of unincorporated King County?* |
| NumberofChildren | *integer* | *The number of children under age 18 who live with the Kinship Caregiver* |
| RelationshipToRecipientCode, RelationshipToRecipient | *integer* | *What is the relationship of the caregiver to the care recipient?* |
| | 0 | Unkown |
| | 1 | Husband |
| | 2 | Wife |
| | 3 | Son/Son-in-Law |
| | 4 | Daughter/Daughter-in-Law |
| | 5 | Grandparent (Kinship) |
| | 6 | Other Relative |
| | 7 | Other Non-Relative |
| | 8 | Other Elderly Relative Caregiver (Kinship) |
| | 9 | Other Elderly Non-Relative Caregiver (Kinship) |
| Kinship | *text* | *Is this a grandparent or older adult caring for a child(ren) under age 18?* |
| | u | Unkown |
| | y | Yes |
| | n | No |
| Veteran | *text* | *Is the client a Veteran?* |
| | u | Unkown |
| | y | Yes |
| | n | No |
| **ADLs** | *text* | *Does the client need help with the following Activiites of Daily Living (ADLs)?* |
| Eating | y | Yes |
| | n | No |
| Toileting | y | Yes |
| | n | No |
| Walking | y | Yes |
| | n | No |
| GettingPlaces | y | Yes |

| | | |
|---|---|---|
| | n | No |
| Transferring | y | Yes |
| | n | No |
| Dressing | y | Yes |
| | n | No |
| Bathing | y | Yes |
| | n | No |
| MedicalManagement | y | Yes |
| | n | No |
| **IADLS** | *text* | *Does the client need help with the following Insturmental Activities of Daily Living (IADLs)?* |
| Cooking | y | Yes |
| | n | No |
| Shopping | y | Yes |
| | n | No |
| Chores | y | Yes |
| | n | No |
| Driving | y | Yes |
| | n | No |
| HeavyHousework | y | Yes |
| | n | No |
| Phoning | y | Yes |
| | n | No |
| MoneyManagement | y | Yes |
| | n | No |
| DivisionID | *integer* | *Code for HSD division for contract* |
| ServiceMonth | *integer* | *Month that service was provided* |
| Serviceyear | *integer* | *Year that service was provided* |
| AgencyID | *integer* | *Unique identifier for agency providing service* |
| SiteID | *integer* | *Unique identifier for site of agency* |
| ServiceAreaID, ServiceArea | *integer* | *What area of service was provided for the client?* |
| | -1 | Unspecified |
| | 8 | Transportation |
| | 11 | Nutrition |
| | 15 | Adult Day Health |
| | 17 | Mental Health |
| | 19 | Case Management |
| | 44 | Family Caregiver Support |
| | 47 | Information & Assistance |
| | 83 | Elder Abuse |
| | 123 | Health Promotion |
| | 138 | CFF - Client Flexible Funds |

| | | |
|---|---|---|
| ServiceTypeID | *integer* | *What type of service was provided for the client? Each integer value maps to a specific service type from a provider. Each ServiceTypeID aligns with a specific ServiceArea* |
| UnitsProvided | *number* | *Number of units provided in service to client* |
| UnitsProvidedType | *text* | *Type of unit for service provided (e.g. hours, meals, session)* |
| ContractID | *integer* | *Unique identifier for specific HSD contract associated with record* |

***Appendix Figure B -*** All Binary/Dummy Variables Explored Visualized by Presence or Absence (e.g., respectively 1 = Yes, and 0 = No)



## Activities of Daily Living (ADLs) help needed by users.