Integrity Client Valuation Report

Summary

Employing several data mining models and a variety of dwelling data, mortgage data, and demographic data, the financial advising group, Integrity, has tasked us to predict which clients are *high value* (annual income >$150K). With an average lifetime value of $60K, the yield of correctly identifying a high value client is $1200, while a loss of $600 is attached to misidentifying a low value client as high value. The goal is to create a model using a total of 20,000 households that provides the most financially beneficial predictions and maximizes profit.

Data Preparation

The first step included changing the variable types to the correct format. The **continuous** variables included Bedroom, Water, Zestimate, Nchildren, Nperson, Vehicles, Rooms. The **ordinal** variables were Lot, Built, Internet, Move, Over60, Over65, Workers. All other variables were set to **nominal**.

Then, we looked at the outliers by examining the distributions and explored outliers. The ones we flagged were Water, Zestimate, and SecMtg, but ultimately decided that SecMtg was the only one that had real outliers. We recoded the SecMtg data that did not fall in the range of 0-4 to missing.

After that, we looked at the continuous variables that had missing data, which were Bedroom, Rooms, Npersons, and Zestimate. We decided to use informative missing for all the variables except for Npersons, which we used imputation. For Npersons, we use imputation because correlation is high with nChildren and it was not missing at the same time as other missing variables. The missing data for Bedroom, Rooms, and Zestimate seemed less random, as there were 690 cases where they were all missing together. Furthermore, we decided to use informative missing for bedroom and rooms because they are highly correlated but have both have missing data. Zestimate also has high correlation with rooms and bedrooms.

The nominal and ordinal variables that required recoding were Lot, SecMtgStat, FamEmp, ChildAge, Workers, WkExp, and WkStatus. We recoded these variables to missing.

We decided not to bin any variables as each level provided meaningful information that would not add more value if binned.

Modeling

Regression

The logistic regression performed relatively poorly in comparison to the other models that ran this data. It would yield a lower profit compared to other models, as it is not designed to function

with profit maximizing in mind. Additionally, the cutoff is fixed and not profit-driven, so it does not adjust based on the data given. The limitation exists as it relies on linear assumptions and does not have the capacity to capture complex interactions like some of the other models.

Decision Tree (Partition)
The decision tree models performed reasonably well on this dataset and showed strong generalization. The validation R^2 values were very close to the training R^2 values (around 0.39-0.40), suggesting that the models were not overfitting. Validation misclassification rates were consistently near 9-10%, indicating solid classification accuracy.

While the decision trees were easy to interpret and captured nonlinear relationships, they tended to under predict high value households. As a result, the profit generated by these models was moderate compared to ensemble methods. Adding a profit matrix improved performance slightly, but decision trees still underperformed relative to more advanced approaches.

Bootstrap Forest (Without Tuning Table)
The bootstrap forest models improved upon the decision tree results by increasing stability and predictive accuracy. These models achieved higher validation R^2 values, around 0.54, with validation misclassification rates around 8-9%.

In terms of financial performance, the bootstrap forest generated meaningfully higher profits than the decision tree models, showing that averaging across many trees helped better identify high-value households. Although not the top performing model overall, the bootstrap forest provided a strong balance between accuracy, consistency, and profit.

Boosted Tree (Without Tuning Table)
The boosted tree models performed well and consistently outperformed both decision trees and bootstrap forests. Validation R^2 values were around 0.43, and misclassification rates were generally below 9%.

Most importantly, boosted trees produced significantly higher profits, ranging from around $310,000-$317,000. This indicates that the boosting process was effective at improving predictions for harder to classify households and aligning the model with the profit objective. While less interpretable than simple trees, boosted trees proved to be one of the stronger non-tuned model in this analysis.

XGBoost (Without Tuning Table)
XGBoost was one of the strongest-performing modeling technique in this analysis. Across multiple runs using JMP's recommended default settings, the model consistently generated substantially higher profits than all other previous approaches, with an average profit of approximately $538,950. To further test model sensitivity, we decreased the learning rate to 0.10 and increased the number of iterations to 300, allowing the model to learn more gradually while

maintaining strong performance. Overall, XGBoost demonstrated superior ability to balance predictive accuracy and economic outcomes, making it one of the most effective model for this dataset.

XGBoost (With Tuning Table)
We ran at total of six XGBoost (with Tuning Table) models and all of them predicted a higher profit than the previous models tested. The models varied from 20 to 100 iterations to determine the optimal balance between model complexity and performance. Overall, XGBoost performed well because it can capture non-linear relationships and complex interactions within the data, which allowed it to better align predictions with profit maximization.

The best-performing model achieved a validation R^2 of 0.4244 and generated a predicted profit of $560,400. Unlike simpler models, XGBoost is more flexible and data-driven, enabling it to adapt to patterns in the data rather than relying on strict linear assumptions. This makes it particularly effective for profit-focused modeling despite being more complex and computationally intensive.

Neural Net
The neural network performed reasonably well relative to other models. It achieved moderately high R^2 values, indicating a decent fit to the data. However, despite its predictive strength, the model generated relatively low profit compared to tree-based methods, particularly XGBoost. This suggests that while the neural network captured overall patterns, it was less effective at optimizing the profit-driven classification objective.

Neural Net (With Tuning Table)
With the tuning table, we were able to home in on specific hyperparameters that maximized profit using K-fold cross validation. The tuning process focused on trying different configurations and identified the model that yielded the highest profit. By running multiple models and changing the configurations, such as validation columns, K-folds, One vs Two Hidden Layers, and number of runs, we successfully achieved a model that performed better than all of our other models.

Model Evaluation Table

Technique	# of Models Run	Comment on the efficacy of this technique on this data set
Regression Based Models	1	In the logistic regression, the validation misclassification rate was .1409 or 14.09%. The validation R^2 is .3419. This is not the greatest model to run.

Decision Tree (Partition)	1	One pruned tree model was run. The validation R^2 is 0.397, which is nearly identical to the training R^2 of 0.425, indicating that the model generalizes well and is not overfitted. The validation misclassification rate is 0.0984 or 9.84%
Bootstrap Forest (Without Tuning Table)	1	The Bootstrap forest ran and yielded a validation R^2 of .5363. and a training R^2 of .6311. The misclassification of the validation data was .0874 or 8.74%, which was better than the decision tree. The profit was 397(1200) + 226(-600) = $340800.
Boosted Tree (Without Tuning Table)	3	1st boosted tree: The boosted tree produced a validation R^2 of 0.4324 and a misclassification rate of 0.0904, compared to a training misclassification rate of 0.0756. The validation profit was calculated as (220 × –600) + (374 × 1,200) = $316,800. 2nd boosted tree: - Validation misclassification: .0896 vs .0755 training - Validation r2 of .4320 - Validation Profit of $315,600 3rd boosted tree - Validation misclassification: .0880 vs .07 training - Validation R2: .4299 - Validation profit: $310,800 This model performed well, but does not result in the highest profit.
Neural Net	4	1st neural net: - Set number of tours=3 - Validation R^2 is .5233 - Training R^2 is .5458 - The misclassification rate is 0.0874 - The profit is 274 (1200) + 101 (-600) = $268,200. 2nd neural net: - Set number of tours=4 - Validation R^2 is .5220 - Training R^2 is .5407 - The misclassification rate is 0.0898 - The profit is 256 (1200) + 95 (-600) = $250,200 3rd neural net: - Set number of tours=5 - Validation R^2 is .5236 - Training R^2 is .5407

		- The misclassification rate is 0.0898 - The profit is 260 (1200) + 97 (-600) = $253,800 4th neural net: - Set number of tours = 9 - Validation R^2 is .5245 - Training R^2 is .5430 - The misclassification rate is 0.089 - The profit is 266 (1200) + 101 (-600) = $258,600
XG Boost (Without Tuning Table	4	**XGBoost 1:** Validation R^2 = 0.2437, Training R^2 = 0.2956, Misclassification Rate = 0.0769, Profit = $764 \times 1{,}200 + 626 \times -600 = \$541{,}200$. **XGBoost 2:** Validation R^2 = 0.4178, Validation Misclassification Rate = 0.0961, Validation Profit = $551,400. **XGBoost 3:** Validation R^2 = 0.4176, Validation Misclassification Rate = 0.1000, Validation Profit = $528,000. **XGBoost 4:** Validation R^2 = 0.4175, Validation Misclassification Rate = 0.0996, Validation Profit = $535,200. This model performs well with increased profits.

Technique	Total Number of Runs	# of Tables Created	Comment on the efficacy of this technique on this data set
XGBoost (With Tuning Table)	6	450	Our 1st XGBoost (w/ tuning table) had a validation R^2 of -0.008 and training R^2 0f -0.003. The misclassification rate is .1083. The profit is 783(1200) + 978(-600) = $352,800. 2nd xgboost (w/tuning): - Set min to 20, max to 100, number of design points to 10, number of inner folds to 2 - Validation R^2 is .4257 - Training R^2 is .4553 - Misclassification rate is .0892 - Profit is 750(1200) + 524(-600) = $549,600 3rd xgboost (w/tuning): - Set min to 40, max to 100, iterations 30, number of design points to 12, number of inner folds to 2 - Validation R^2 is .4244

			- Training R^2 is .4466 - Misclassification rate is .0898 - Profit is 728(1200) + 522(-600) = $560,400 4th xgboost (w/tuning): - Set min to 35, max to 100, number of design points to 7, number of inner folds to 2 - Validation R^2 is .4204 - Training R^2 is .4931 - Misclassification rate is .0766 - Profit is 706(1200) + 518(-600) = $536,400 5th xgboost (w/tuning): - Set min to 40, max to 100, iterations 50, number of design points to 12, number of inner folds to 2 - Validation R^2 is .4252 - Training R^2 is .4771 - Misclassification rate is .0897 - Profit is 722(1200) + 517(-600) = $556,200 6th xgboost (w/tuning): - Set min to 40, max to 100, iterations 50, number of design points to 12, number of folds to 5 - Validation R^2 is .4227 - Training R^2 is .4682 - Misclassification rate is .0910 - Profit is 717(1200) + 510(-600) = $554,400
Neural Net (W/ Tuning Table)	4	135	1st Neural Net - 25 runs - One Hidden Layer - Validation + informative missing - Selected Neural Net 10 o Validation misclassification Rate: 0.0932 o Validation R^2 = .4995 o Validation Profit is 128(-600) + 272(1200) = $249600 2nd Neural Net - 50 runs - Selected Neural Net 30 o Validation misclassification rate: .0996 o Validation R^2 = .4521 o Training R^2 = .5787 o Validation Profit: 167(-600) + 279(1200) = 234,000 3rd Neural Net

			- Crashes JMP when Validation Column with Two Hidden Layers, 0 min 9 max TanH, Linear, Gaussian 4th Neural Net with **K-folds** - 60 runs - TanH and Gaussian / Two Hidden Layers - K-Folds - Row 1 o 18000 Training, 2000 Validation o Mean = 84.54 o Misclassification Rate = .0729 o Validation R^2 = .6585 o Training R^2 = .6547 o Training Profit (18k) = 423(-600) + 1386(1200) = 1,409,400 o Validation profit (2k) = 54(-600)+157(1200) = 156,000 o Total Profit = 1,565,400 - Row 2 o Misclassification Rate = .073 o Validation R^2 = .6261 o Training R^2 = .6128 o Training Profit (18k) = 426(-600) + 1249(1200) = 1,243,000 o Validation profit (2k) = 46(-600) + 145(1200) = 146,400 o Total Profit = 1,389,400

Model Selection

The model we selected to make our best predictions was the *Neural Network with Tuning Table.* With a .073 misclassification rate, it correctly predicted 18,541 of the 20,000 records in the combined batch 1 and 2 data. This specific model was made with 60 runs, 10 K-Folds, Informative Missing, Two Hidden Layers, Squared penalty method, Transform Covariates set to Include in DOE and 0 min and 9 max in TanH and Gaussian, choosing to forgo Linear.

After the model ran, we compared the top 20 results and looked at the mean and sum values of "Actual Profit for HiValue #)." The model we eventually selected, NN Model 2, had the highest mean at 84.4. We looked more into this model, and it also provided us with the highest profit we had seen yet at $1,690,800. We did more checks to see if this model was a viable option, like changing the data and validating it was as effective as we'd hoped. By applying our new model to batch 2 data, we found that it maximized profit at $828,600 compared to $568,200 achieved with our previous model.

We decided to use K-Folds over Validation column as our initial runs using validation ran poorly, even compared to XGBoost with tuning. Despite changing the parameters for the NN with validation column, it still only gave a profit of ~$250,000 with the 5,000 validation rows. However, with the validation we only tried one hidden layer opposed to two with the K-Folds. The two hidden layers were not a viable option for validation column as it would not run well and crash the program.

Actionable Insights

a. Integrity's goal is to identify households that are likely to be high-value (homeowners with income over $150K) and, ideally, have excess cash available to invest. To build on our work, Integrity should prioritize collecting data that better captures wealth proxies and spending capacity at the household level.

   First we suggest Integrity collects zip codes, rather than relying on state. ZIP code would be far more informative than state because wealth and income vary dramatically within a state, and ZIP-level location is a strong proxy for property values and affluence.

   Next we suggest they collect vehicle type rather than just number of vehicle. The current Vehicles variable only measures quantity (0–6+), but the type of vehicle (luxury vs. economy, new vs. older) is more likely to signal discretionary spending and wealth. We do suggest keeping number of vehicles as this turned out to be an important factor in many of our models.

   Lastly we suggest adding an education-related spending factor. If Integrity can obtain third-party indicators related to schooling (private school enrollment, tuition payments, or neighborhood school type), that could help identify affluent families and long-term client potential.

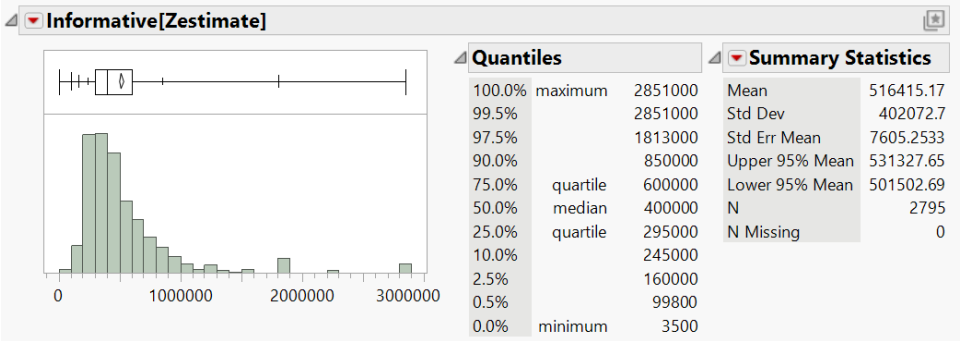
   Integrity should avoid spending time collecting additional broad geographic fields (like region/state beyond what's already included), since ZIP code would dominate that information. They also don't likely need to collect both number of bedrooms and rooms, as these factors are highly correlated and one can tell the full story. Furthermore, the Zestimate of the house gives enough detail on home value. Another variable we found unnecessary to collect is language, as it did not strongly influence any of our models.

b. We recommend that Integrity prioritize obtaining higher-quality data for Zestimate, as property value is a strong indicator of household wealth. Zestimate was an important factor in all of our best models, especially XGBoost. The key issue with the current Zestimate variable is the amount of missing data. While we imputed these values in our analysis, access to more complete and accurate property valuations would improve the

model's ability to predict high-value households, since true values are more informative than imputed estimates.

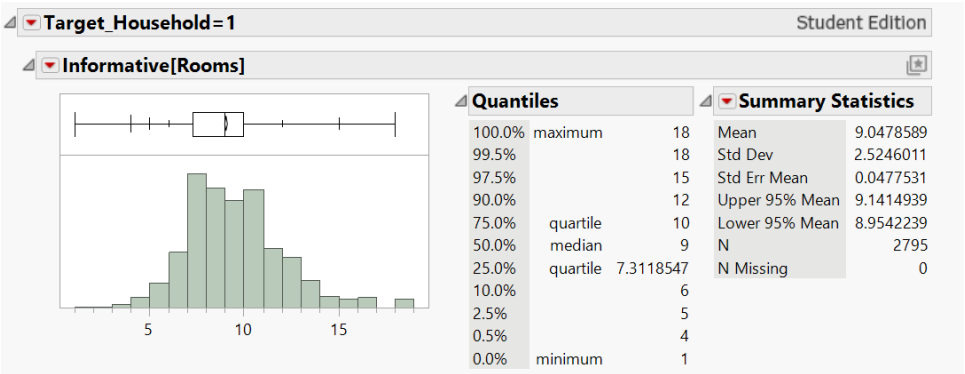
c. To identify the households Integrity should pursue more aggressively, we used our final neural network model and focused on expected profitability rather than prediction accuracy alone. A household was labeled as a target if the model predicted a positive expected profit, meaning the expected benefit of marketing to that household outweighed the potential cost. This profit-based approach aligns directly with Integrity's high touch marketing strategy. By examining the characteristics of these target households, we can determine the type of household the firm should prioritize more aggressively.

Chart 1: Zestimate by target group



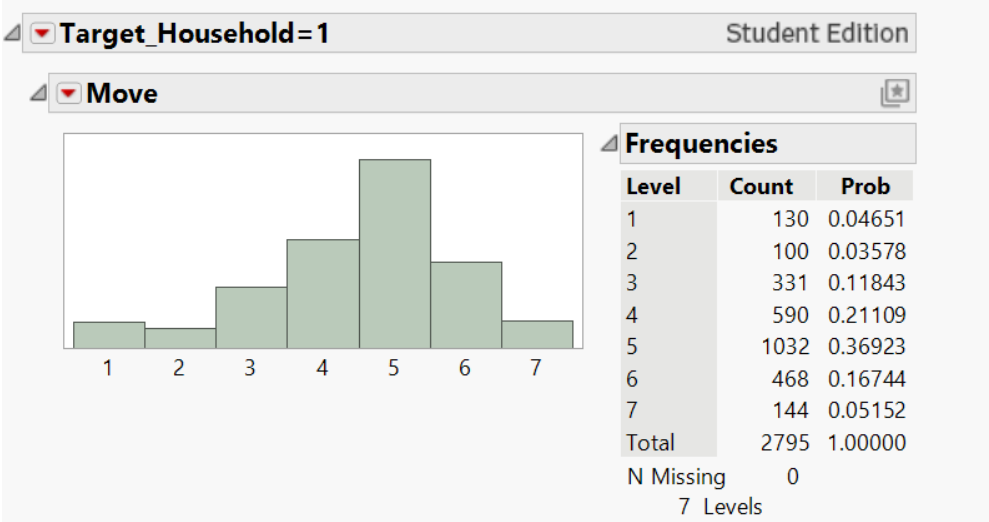
This data shows that targeted households have very high property values. The median Zestimate is approximately $400,000, the upper quartile reaches $600,000+ with a long right tail, and very few targeted households have low home values. Zestimate is one of the strongest proxies for household wealth in the dataset. The neural clearly favors households with higher valued homes because they are more likely to have excess assets and long-term advisory needs.

Chart 2: Home size (rooms) by target group



This data shows that the mean number of rooms is around 9, the median is 9 rooms, the interquartile range is roughly 7-10 rooms, and targeted households tend to live in larger than average homes. Home size reinforces the Zestimate results by capturing physical scale and lifestyle. Larger homes typically correlate with higher income, wealth, and discretionary spending capacity. The neural network consistently selects households living in larger homes.

Chart 3: Length of residence (Move) by target group



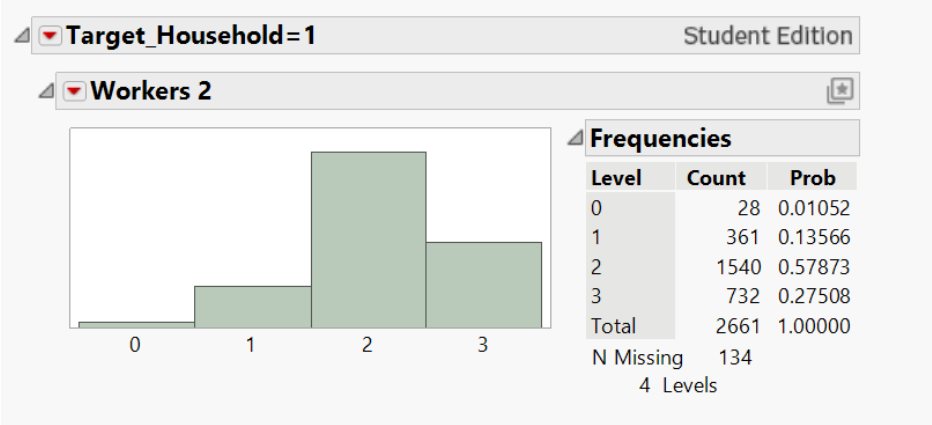
The data shows that over 36% of targeted households fall into Move =5, more than 60% have lived in their home 5 years or longer, and very few targets have short residence tenure. Longer residence tenure signals financial stability, established homeownership, and lower mobility. Stable, long-term homeowners are significantly more likely to be profitable targets.

Taken together, these results indicate that the households Integrity should target most aggressively are affluent, established homeowners with high property values, larger homes, and long-term residence stability. These households consistently generate positive expected profit in the neural network model and exhibit clear indicators of wealth and financial maturity. Focusing marketing resources on this segment allows Integrity to maximize the return on its costly outreach efforts while prioritizing households that are most likely to benefit from and engage with long term financial advisory services.

d. In addition to targeting affluent, estimated homeworkers, Integrity should also pursue working professional households that exhibit strong earning potential and active lifestyles. Using the neural network's profit-based target classification, we analyzed

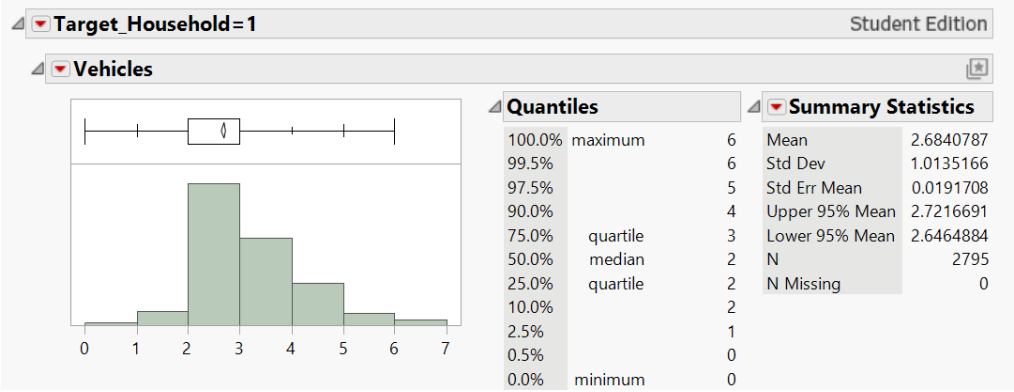
characteristics of profitable households that differ from the core segment identified in Part C. These households may not be the most established homeowners, but they still generate positive expected profit and represent a valuable secondary target segment.

Chart 1: Workers by Target Household



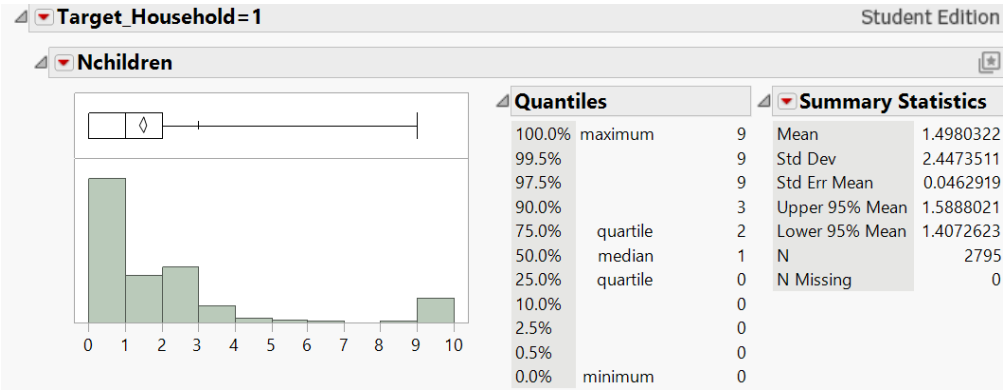
This chart shows that the majority of targeted households have two or more workers, with nearly 58% of target households having two workers and an additional 28% having three or more workers. Very few targeted households have no workers. This indicates that dual-income and multi-worker households are strongly associated with profitability, as they signal higher and more stable earning capacity.

Chart 2: Vehicles by Target Household



The vehicles distribution further supports this profile. Targeted households tend to own multiple vehicles, with a median of two vehicles and an average close to three. Multiple vehicles suggest higher discretionary spending and an active lifestyle, both of which align with households that have greater capacity for investment and financial planning services.

Chart 3: Presence of Children

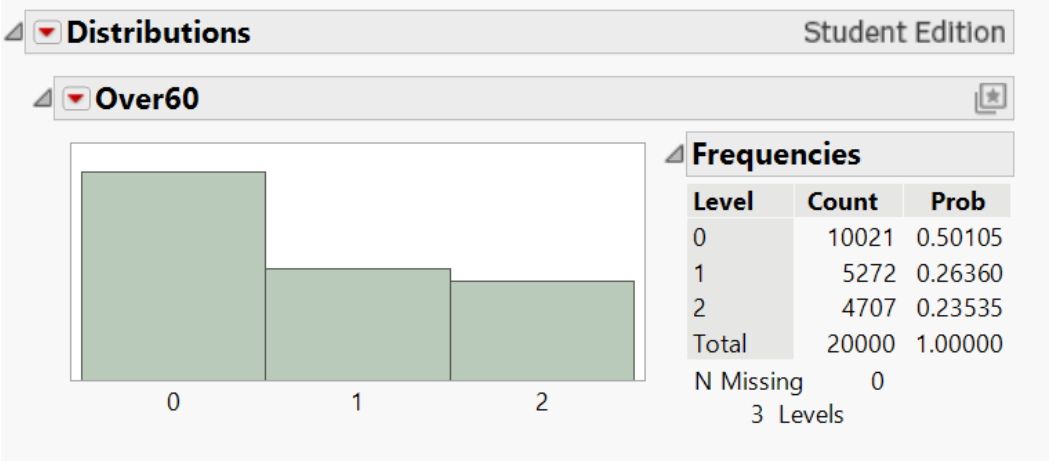


The presence of children data shows that most targeted households have one or more children, with the median number of children equal to one. This suggests that many profitable households are family-oriented and likely facing financial decisions related to education, savings, and long-term planning. These needs make them strong candidates for Integrity's advisory services, even if they differ from the more established homeowners highlighted in Part C.

Together, these charts indicate that a second attractive segment for Integrity consists of working, family-oriented households with multiple earners, multiple vehicles, and children present in the household. While this group may not always exhibit the highest home values or longest residence tenure, their strong income potential and active financial needs make them a profitable and strategically valuable segment. By targeting these households alongside its core affluent homeowner segment, Integrity can expand its client base while maintaining alignment with its profit-maximizing marketing strategy.

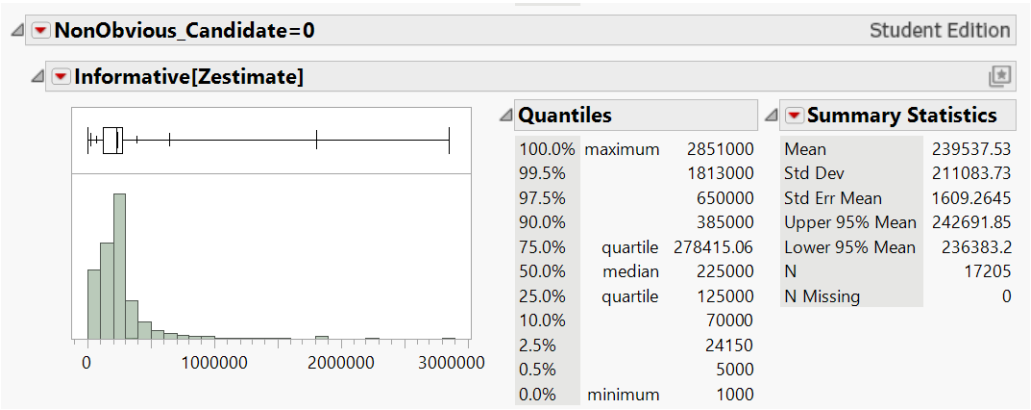
e. To identify a non-obvious but profitable household type, we created an additional column using the neural network's expected profit output rather than traditional wealth measures. This approach allows us to analyze households the model still values economically despite not fitting standard targeting assumptions.

Chart 1: Age Indication (Over 60)



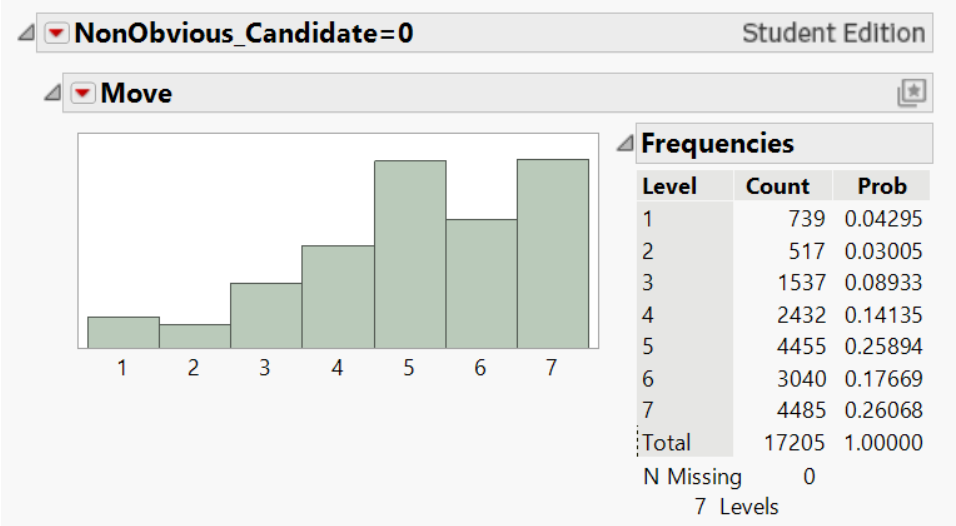
The age distribution shows that a substantial share of households include one or more individuals over the age of 60. Older households are often overlooked in acquisition strategies due to lower labor participation or fewer workers, but they frequently possess accumulated assets and face complex financial planning needs related to retirement, estate planning, and wealth preservation. This makes them strong candidates for advisory services despite not appearing traditionally affluent.

Chart 2: Home Value (Lower Zestimate but still profitable)



The home value distribution indicates that many households fall below high-end Zestimate thresholds, with a median value around $225,000. While these homes are not among the most expensive in the dataset, the neural network still identifies value in this segment. This suggests that home value alone does not fully capture household quality and that profitable opportunities exist beyond conventional wealth cutoffs.

Chart 3: Stability (Move)



The length-of-residence data shows that most households have lived in their homes for five years or more, signaling long-term stability and commitment to homeownership. Stable households are more likely to engage in ongoing financial planning relationships, making them suitable targets for Integrity's advisory model even if they do not exhibit obvious signs of affluence.

Together, these charts suggest that Integrity should consider targeting stable, older households with modest home values and long residence tenure as a non-obvious but valuable segment. While these households may not stand out based on traditional wealth metrics, the neural network's predictions indicate meaningful advisory potential driven by stability, accumulated assets, and long-term planning needs. Expanding outreach to this segment allows Integrity to uncover profitable opportunities that conventional targeting approaches may miss.