# Bedding Bathing & Yonder (BBY)– Household Spend Prediction Model

**Purpose**

To build a prediction model for the household spending for Beddin Bathing & Yonder stores.

**Exploratory Data Analysis**

A graph showing the number of people missing rows

Description automatically generatedAround half of the eighty attributes available for our predictive model were dropped in the preliminary analysis. Some were unnecessary, like names or unrelated to household spending, and some unethical attributes, like religious or political affiliations or even race and donations. For example, if you are a horse owner or not, there is no relation to buying domestic merchandise. Numerical and categorical had missing values and outlier treatments. The treatment data was split into 80% for training and 20% for validation sets. The attributes with missing values were in almost all of the cases. After the first dropping of features, the features with missing values are around half of all the features.

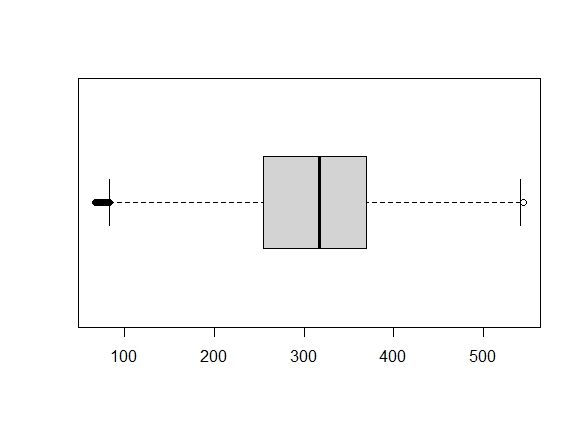
**Numerical Attributes**

The attributes with more than 40% missing values were dropped from the model for the numerical values. Before the missing value and outlier analysis, a correlation between the variables was ruined, only revealing weak to no correlation between them. The missing values treatment has the following rules.

* For insignificant missing percentage [ <2%]: Rows were removed.
* For the rest of the cases [%<40]: Median or mean imputation, depending on their distribution.

For the outlier treatment, the household spending and home values attributes had less than 5% of missing values, while the land value attribute had less than 10%. For household spending, outliers were replaced with the upper limit value of its distribution, capping the values.

A graph with a line and a square

Description automatically generated with medium confidence

Household Spend Boxplot w/Outliers

Household Spend Boxplot w/o Outliers

Capping

In the cases of home values, outliers were left in the data as they could represent natural variations. There are few exclusive areas in the cities where the home values are incredibly high compared to the other areas of the same city, and there are more than 4000 cities in the data.

With the remaining attributes, a correlation was rerun between them, showing only weak relations again with household spending (“yhat”). It was decided to keep the variables with some correlation with the variable to predict, which were only 2, and the variables that correlated with both 2.

A graph showing the results of a performance

Description automatically generated with medium confidence

*Numerical Variables Correlation after missing values and outlier treatment.*

It was surprising at first that the frequency of visits is not a driving factor, but Bedding Bathing & Yonder has an online presence, and people might visit the store to compare prices as most of the entries have had six times visits over the last 12 months. Finally, the numerical variables to be included in the model were ISPSA, Median Education Years, Age and store visit frequency.

**Categorical Values**

A graph showing a number of property type

Description automatically generatedA graph showing the number of property type

Description automatically generatedThe missing percentage threshold for these cases was 70% as some categorical variables with missing values could be assumed to be “No” or lack attributes like “Investor.” Missing value rows were removed for categories with less than 2%. There are some variables with more than 85% of the same value in their histograms, like “Veteran” and others like “city” of “county,” with lots of unique values that had to be dropped. Other cases, like the Occupation Industry, where unknown values were almost 70%, were also dropped to avoid overfitting. Values with less than 5% frequency between the attributes were grouped and categorized in the " other " variable to reduce overfitting.

Grouping

There were other cases like. Finally, the categorical variables for the model would be the Presence of Children, Home Owner Renter, Net Worth, Investor, Education, amount of books bought, Collector in Household, Computer Owner, Gender, state, Dwelling Unit Size, and Property Type.

**Modeling**:

After the preparation, the informative variables were sixteen out of the initial eighty and underwent a treatment to use the categorical values in the models. Four distinct models were iterated to get the best predictor of spending: the linear model with two iterations and the decision tree model with two iterations. The first iteration was done with the variables selected after the EDA, and the second iteration was done with the significant variables that the linear regression showed in its model.

**Evaluation**:

Even though the best Root Median Square Error (RMSE) corresponds to the Decision Tree model under all the variables available, both the best and the second best models were evaluated to see how they behave with the training set, the validation set, and the test set with the following results:

|  |  |  |
| --- | --- | --- |
|  | **RMSE** | |
| **Set** | Decision Tree | Linear Model |
| **Train** | 89.41338 | 93.51813 |
| **Validation** | 91.51989 | 95.16957 |
| **Test** | 89.41338 | 94.13914 |

Both models are consistent across the data sets, but the decision tree has a better RMSE; therefore, the final model for this prediction is the decision tree with all of the variables available after the cleaning and treatment process.

**Final model**

The decision tree model with all of the variables selected after the EDA process:

**A diagram of a number

Description automatically generated**

According to the plot, the occupancy type of the customer as a Dwelling Unit Size variable and the presence of a home computer are the leading attributes to predict the household hold expenditure in Bedding and Bathing, and Yonder has four possible outcomes having a 1-single family dwelling as the customers with potentially more expenditure in Bedding Bathing and Yonder.

**Conclusions**

The model has a relatively high RMSE and a mean absolute percentage error of 43%, meaning that predictions of this model deviate from the actual value of around 43%. Using these data sets, the dwelling unit size and the computer ownership are the mean attributes to predict spending, but this could change if the model is better trained. Bedding Bathing and Yoder should consider getting more complete consumer data; even with variables like network or frequency of visits, they are not strongly related to household spending. Also, the location was too dispersed as the maximum entries with the same state were for Texas, with around 5% of the frequency, followed by California. With the complete state information, better models could be built and better aimed at the company's divisions in those areas.