****

**XGB**



**Course: ALY 6020 Predictive Analytics**

**Instructor: Dr. Marco Montes de Oca**

**Group Members:**

**Deepak Prasad Baskaran**

**Nidhi Shah**

**Nithish Saravanan**

**INTRODUCTION**

Data is collected and used to analyze questions, perform test hypothesis or disapprove theories. According to the cross-industry standard process, there are six phases like data requirement, data collection, data processing, data cleaning, exploratory data analysis, and modeling & algorithms. Data cleaning is the process of preparing data for analysis by removing or modifying data set which has incorrect, incomplete or outliers and data cleaning is not simply about deleting information to make space for new data, but rather finding a way to maximize a data set’s quality without necessarily deleting the information. There are several methods for identifying and managing missing value and outliers in the data set. Once data is cleaned, using a variety of techniques referred as exploratory data analysis is the critical process of performing a preliminary investigation on data to find patterns or insights and to check assumption with help descriptive statistics and graphical representation.

Models are key to predicting the outcomes to business decisions. Modeling involves selecting the right data sets, algorithms and variables with right techniques to format data for a particular business problem. An analytical model predicts or classifies data values by essentially drawing a line through data points when applied to test data sets, a model can predict the outcomes based on historical patterns. Model building journey follows key components like creating hypothesis, loading and transforming data, identifying features, choosing the right model and evaluating the model. Optimizing the model for predicting the target variable for more accurately and improving the model performance. Mode selection refers to the problem of selecting a best model from the set models for purpose of decision making or optimization under certainty. Methods like data transformation, exploratory analysis and model specification will assist for choosing the set of candidate models. Model selection is depended on parameter like Akaike information criterion (AIC), Bayes factor and/or the Bayesian information criterion, likelihood-ratio test, stepwise regression, false discovery rate, cross validation and etc.

Mostly commonly AIC and Bayes factor criteria are used for model selection. The Accuracy of the predictive model can be boosted in two ways, either by embracing feature engineering or by applying boosting.

**ANALYSIS**

**ABOUT THE DATASET**

For this project, we have used a dataset of Inter-American Development Bank which consists of socio-economic conditions of Costa Rican Households such as education background, family details, home, electricity details, etc. This data set is given by Inter-American Development Bank on Kaggle. By analyzing socio-economic conditions of household’s characteristics, government can identify which households have the highest need for social welfare assistance. By accurately predicting the household poverty level we can help other countries beyond Costa Rica for assessing the social need. This data set contains 143 variables like home rent, house conditions, education details, household information, etc. and it contains 9557 rows of data. We need to predict the poverty level (Target variable) of household-based information. The target variable is an ordinal variable indicating groups of income level i.e.

1= extreme poverty

2=moderate poverty

3= vulnerable households

4=non-vulnerable households.

For this project, we have used statistical analysis tool R for analysis and visualizing insights.

**DATA CLEANING**

For handling the missing values, we have used two methods based on variables.

1. Replacing missing values with Average of column or zero.

2. Removing variable from the data set which has more missing values.

Next step is handling the missing values, from our analysis for which we have removed the years behind the school variable which is having more than 80% of missing values. For significant variables like monthly rent, average adult education and square mean of adult education, we choose to replace the missing values with the mean of a column and for the variable ‘number of tablets household’, we choose to replace missing values with zero.

We have used Z- score value to identify the outliers in data i.e. values which differ by more than 3 interquartile range are probably outliers. So, we found that around 159 outliers in data.

As extreme values of monthly rent may impact the prediction of the poverty level, handling the outliers is very important, there are several methods available for handling the outliers.

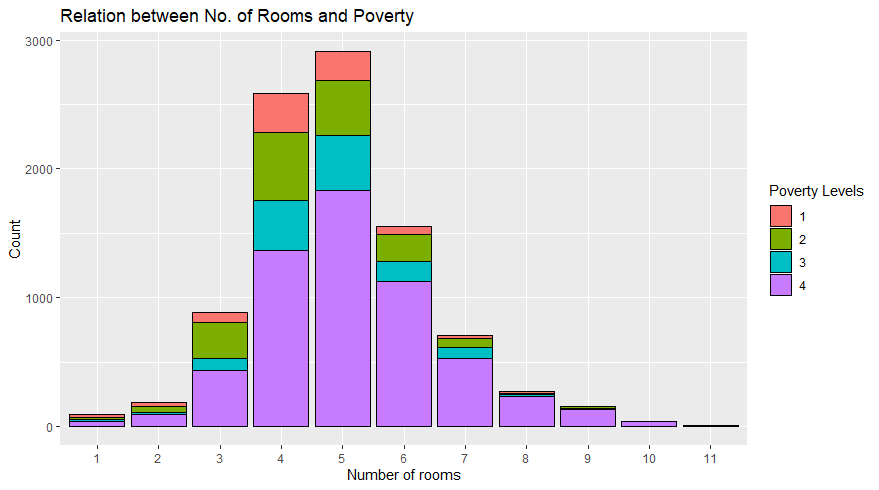
1. Removing the outliers.

2. Consider outliers and rest of the data separately.

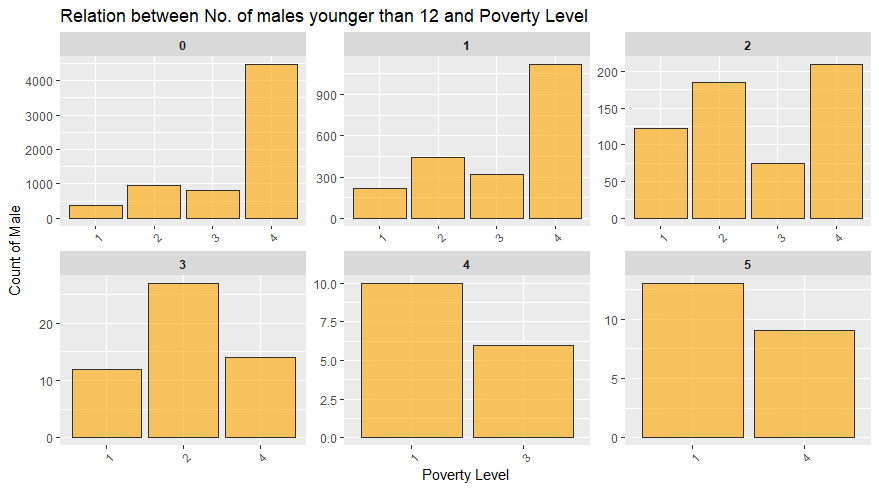
3. Removing and replacing with NA.

We have analyzed the outlier separately and found that all 143 outliers fall under the category of Non-Vulnerable households which might skew the outcome. So, we have removed all the 143 from the data set.

**EXPLOARTORY DATA ANALYSIS**

****

In non-vulnerable households, the highest number of rooms is 5. Whereas extreme and moderate households have maximum rooms of 4.

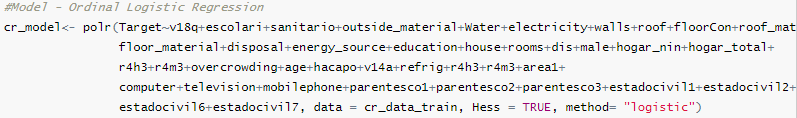


Number of males younger than 12 are high in extreme poverty and non-vulnerable households

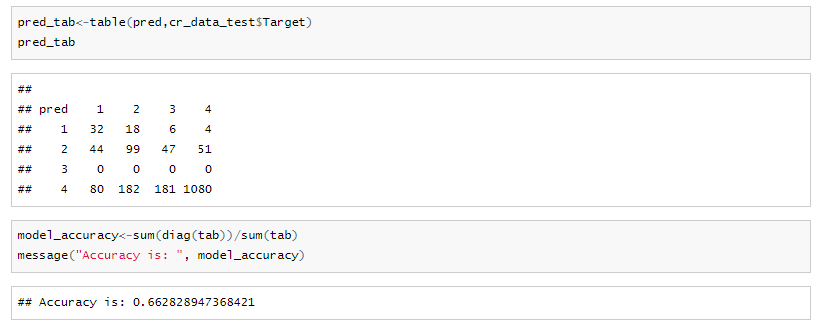
compared to other poverty designations.

**MODEL IMPLEMENTATION**

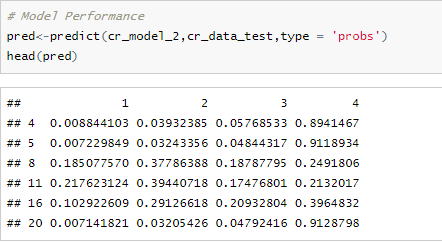
We have used ordinal logistic regression for building the model using ‘polr’ function from ‘MASS’ library by summarizing the model summary using ‘summary ()’ function. We will train the logistic model with the help of training data set.

 From model summary, we get information about

1. The regression coefficients with their values, standard error, Z value and level of significance.
2. AIC and other values, which are used in comparing the model performance



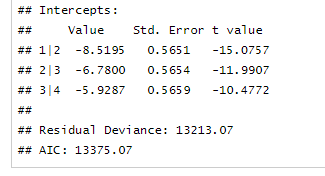
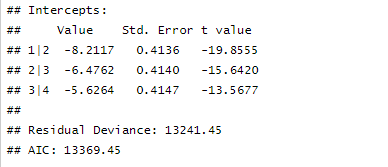
From the confusion matrix we can find that elements in the diagonal are correctly predicted in the model. We can find the accuracy of the model by using diagonal elements. Accuracy of the model is 66%.



To improve the model performance, we have used label encoding by assigning unique value in the feature column and we converted it as ordered variable. To improve the further analysis, we will use feature engineering by removing non-significant variables by iterating model with different combination of variables or by using principle component analysis.

**MODEL SELECTION**

Next Step is model selection, we have combination of variables and AIC values for selecting the best model. Model fit analysis refers to an examination of whether the statistical model employed in an application adequately explains the important features of the data set at hand. The Akaike Information Criterion (AIC) is a way of selecting a model from a set of models. We have used stepwise regression and found that

   
 Model 1 Model 2

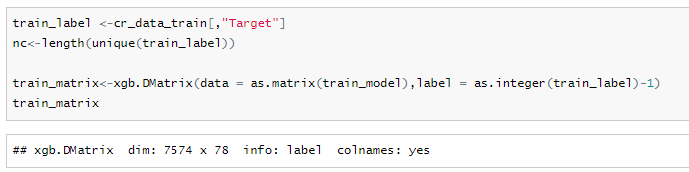
The model with lowest AIC value i.e. 13369.45 is best fit model. From the model selection, we have selected the variables and used to build the new model to compare with old model. We have built new ordinal logistic regression model by using the features selected.

Comparison between two model we didn’t observe much improvement in performance and also observed that performance deviance compared to old model.

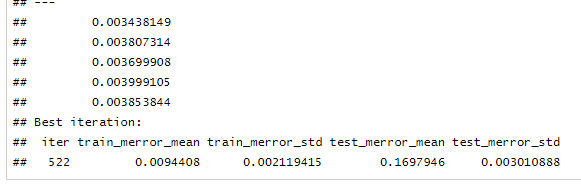
**MODEL OPTIMIZATION**

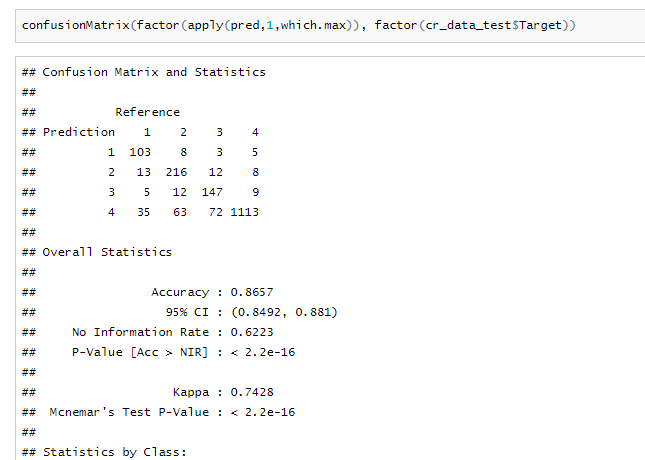
So next step is optimizing model, we have used gradient boosting algorithm. It’s an implementation of gradient boosted decision trees designed for performance and speed. We have used eXtreme Gradient Boosting Package in R to build the model. Boosting is an ensemble technique where new models are added to correct the errors made by existing models. Models are added sequentially until no further improvements can be made. Gradient boosting is an approach where new models are created that predicts the residuals or errors of prior models and then added together to make the final prediction. It uses gradient descent algorithm to minimize the loss when adding new models and due to making only small incremental improvements with each model in the ensemble, this allows us to stop the learning process as soon as overfitting has been detected.

We have created training dataset and testing dataset-based probability of 80% and 20%. First step is creating the matrix for training and testing data sets.so it creates the sparse matrix where numeri variables are unchanged, but it creates the dummy variables for categorical variables.



Next step is creating the parameters for model by using list function like number of classes, training rate and created tuning parameters for model to reduce error and overfitting in the model by using the below functions. In our model we have used cross validation method, to find the ideal time to stop the training model when the validation error decreases and starts to stabiles before it starts increasing due to overfitting and found that 522 iterations is optimum number of rounds for model where test error is minimum and no overfitting in the training model.





From the above model, we found that accuracy of 86 % from XGBoosting model, which means we are able identify 86 % accurately which household belongs to which category of poverty level and we will recommend this model to inter-American development bank to identify which household have the highest need for social assistance.

**View feature importance or influence from the learnt model:**

****

From the above graph, interestingly number of children between the 0 to 19 years-old plays important role for predicting the level poverty along with overcrowding variable. This seems obvious that a greater number of people staying single apartment is correlated with poverty for different reasons. A lot of relative features came up, which is makes interesting inference that living conditions of people and basic amenities for households is actually very important. In addition to obvious variables like level of education, age.

**CONCLUSION**

Based on the accuracy rate, sensitivity and specificity, we can propose this model to inter-American development bank which will help them to analyze socio-economic conditions on household’s characteristics and governments can identify which households have the highest need for social welfare assistance. They can allocate funds for such people who are below poverty line. This solution will also help other countries beyond the Costa Rican for assessing the social need.

**REFERENCE**

* Data Analysis & Exploratory Data Analysis (EDA). (n.d.). Retrieved March 3, 2019, from <https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/data-analysis/>
* Costa Rican Household Poverty Level Prediction. (n.d.). Retrieved June 1, 2019, from

<https://www.kaggle.com/c/costa-rican-household-poverty-prediction/data>.

**APPENDIX**