Project

"Prediction for Costa Rican Households Poverty Level"

SUMMARY

An analytical model predicts or classifies data values by essentially drawing a line through data points when applied to test data sets, the model can predict the outcomes based on historical patterns. 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.

This data set is given by Inter-American Development Bank on Kaggle. By analyzing socio-economic conditions of household's characteristics, governments 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 the 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.

We have used ordinal logistic regression for building the model using 'polr' function from 'MASS' library by summarizing the model summary using 'summary ()' function. 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 reduce the number of variables in the model. So, we have reduced the number of variables in the dataset by transforming data in categorical data for improving the performance of model. 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 the model with the lowest AIC value i.e. 13369.45 is best fit model. By tuning the model, we found the optimum number of iterations of 522 where test error is minimum. We found that accuracy of 86 % after XGBoosting in the model.

INTRODUCTION

Data is collected and used to analyze questions, perform test hypothesis or disprove theories. According to the CRISP-DM, there are phases like understanding the business and data, then we prepare the data for modeling and evaluate the model before deployment.

Data cleaning is the process of preparing data for analysis by removing or modifying data set which has incorrect, incomplete or outliers. There are several methods for identifying and managing missing values and outliers in the data set. Once data is cleaned, using a variety of techniques referred as exploratory data analysis 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 predict the outcomes to business decisions. Modeling involves selecting the right algorithm for the data sets and variables for a particular business problem. Model building journey follows key components like creating hypothesis, loading and transforming data, identifying features, choosing the right model and evaluating the model. 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. Optimizing the model for predicting the target variable for more accuracy and improving the model performance.

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.

We have used R programming language for creating, optimizing the model and generating graphical representations of the data.

ANALYSIS

● ABOUT THE DATASET

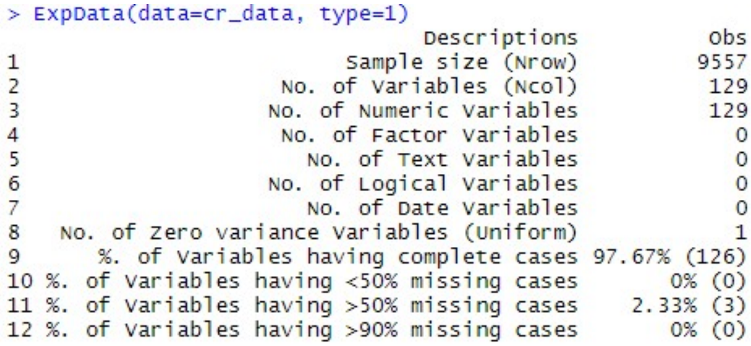
For this project, we used the dataset of Inter-American Development Bank which consists of socio-economic conditions of Costa Rican Households such as education background, family details, home, electricity details, etc,. from Kaggle. By analyzing socio-economic conditions of household's characteristics, 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

We started with identifying the missing values by using R's SmartEDA function. The result is shown in Fig.1



```
> ExpData(data=cr_data, type=1)
                               Descriptions         Obs
1                        Sample size (Nrow)        9557
2                   No. of Variables (Ncol)         129
3                  No. of Numeric Variables         129
4                   No. of Factor Variables           0
5                     No. of Text Variables           0
6                  No. of Logical Variables           0
7                     No. of Date Variables           0
8   No. of Zero variance Variables (Uniform)          1
9      %. of Variables having complete cases 97.67% (126)
10 %. of Variables having <50% missing cases      0% (0)
11 %. of Variables having >50% missing cases   2.33% (3)
12 %. of Variables having >90% missing cases      0% (0)
```

Fig.1 SmartEDA Result

For handling the missing values, we have methods used 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.

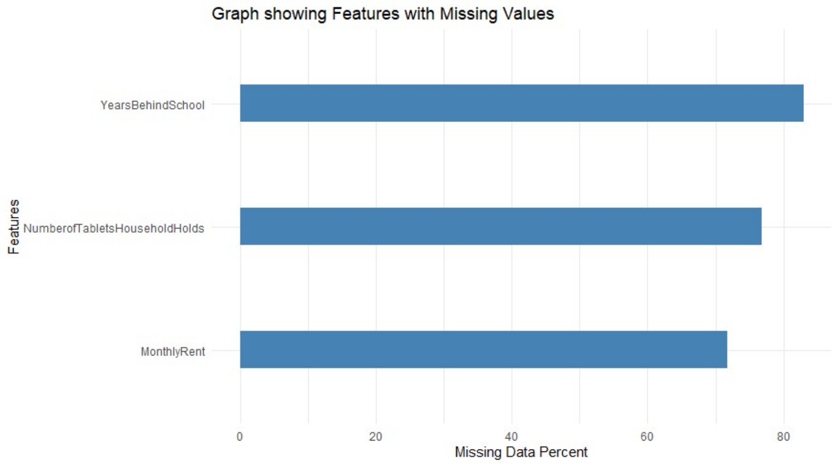


Fig. 2 Missing Values Graph Plot

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 159 outliers fall under the category of Non-Vulnerable households which might skew the outcome. So, we have removed all the 159 from the data set.

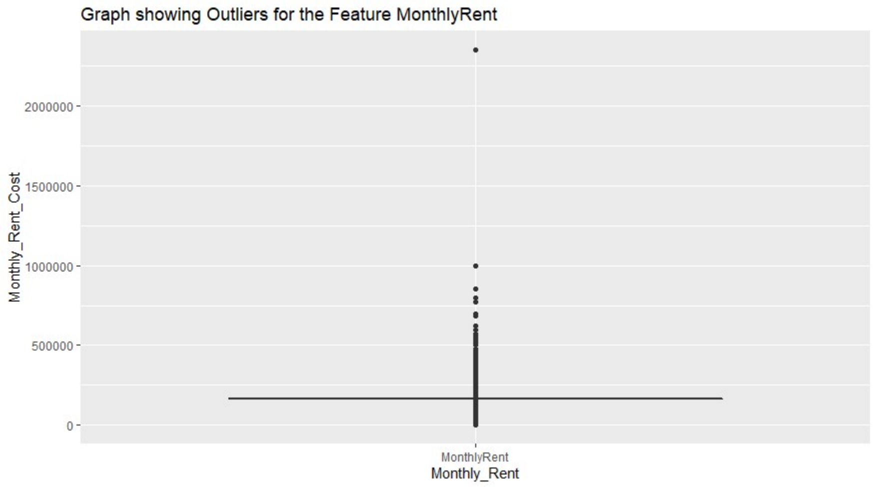
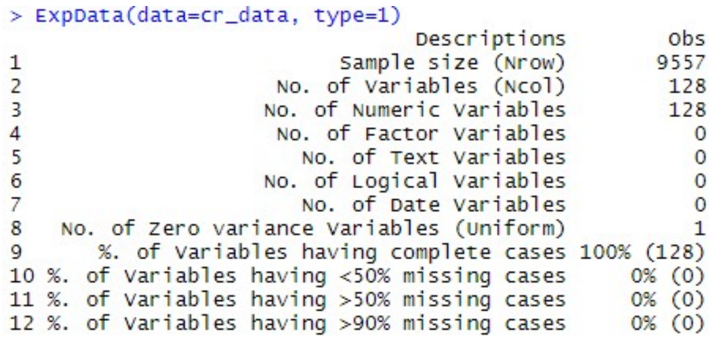


Fig. 3 Outliers Graph Plot for Feature MonthlyRent

Lastly, after receiving the clean data we verified it and started exploration of the data for analysis.



```
> ExpData(data=cr_data, type=1)
                             Descriptions        Obs
1                          Sample size (Nrow)    9557
2                        No. of variables (Ncol)  128
3                      No. of Numeric variables   128
4                       No. of Factor variables     0
5                         No. of Text variables     0
6                      No. of Logical variables     0
7                         No. of Date variables     0
8   No. of Zero variance variables (Uniform)        1
9      %. of variables having complete cases 100% (128)
10 %. of variables having <50% missing cases     0% (0)
11 %. of variables having >50% missing cases     0% (0)
12 %. of variables having >90% missing cases     0% (0)
```

Fig. 4 Verified EDA result with no missing values

- **EXPLORATORY DATA ANALYSIS**

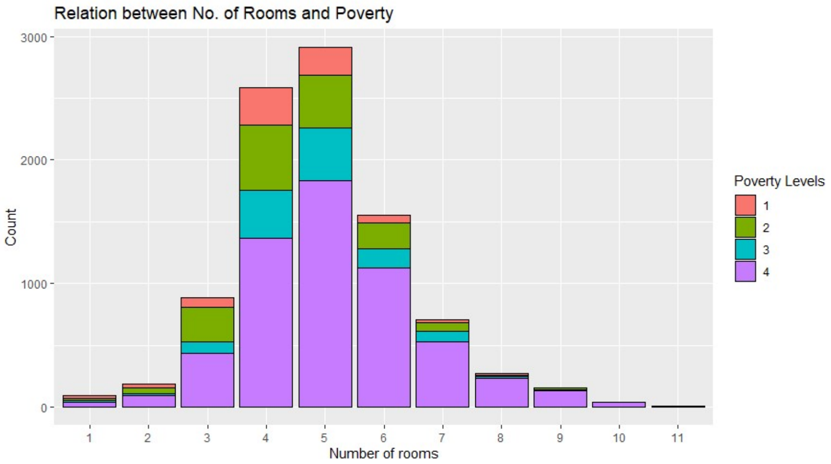


Fig. 5 Graph showing relation between No. of rooms and Poverty Level

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

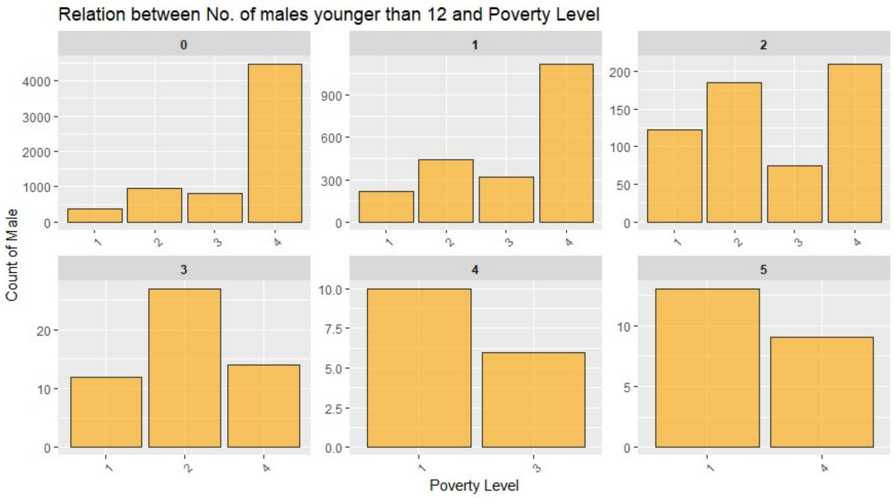
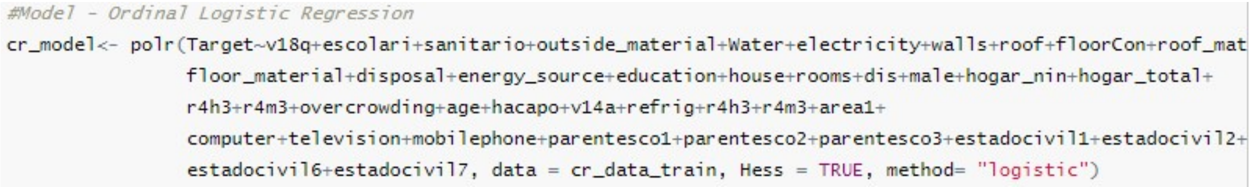


Fig. 6 Representation of feature relation with Poverty Level

Number of males younger than 12 are high in extreme poverty and non-vulnerable households compared to other poverty designations as shown in fig. 6.

- **MODEL IMPLEMENTATION**

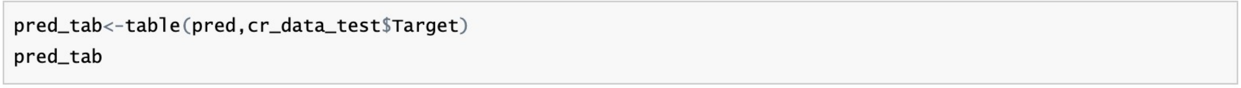
We have used ordinal logistic regression for building the model using ‘polr’ function from ‘MASS’ library. We will train the logistic model with the help of training data set.



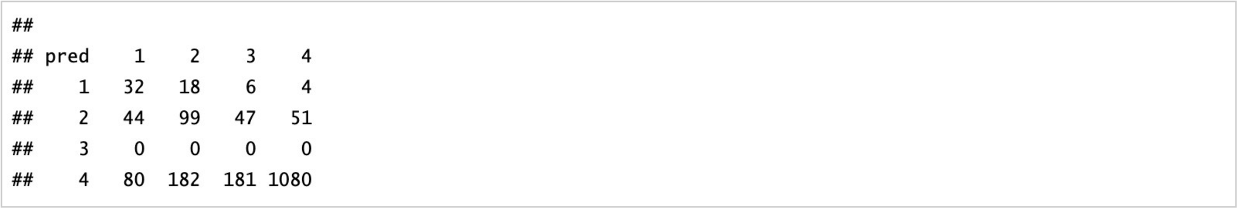
```
#Model - Ordinal Logistic Regression
cr_model<- polr(Target~v18q+escolari+sanitario+outside_material+Water+electricity+walls+roof+floorCon+roof_mat
                floor_material+disposal+energy_source+education+house+rooms+dis+male+hogar_nin+hogar_total+
                r4h3+r4m3+overcrowding+age+hacapo+v14a+refrig+r4h3+r4m3+area1+
                computer+television+mobilephone+parentesco1+parentesco2+parentesco3+estadocivil1+estadocivil2+
                estadocivil6+estadocivil7, data = cr_data_train, Hess = TRUE, method= "logistic")
```

From model summary, we get information about

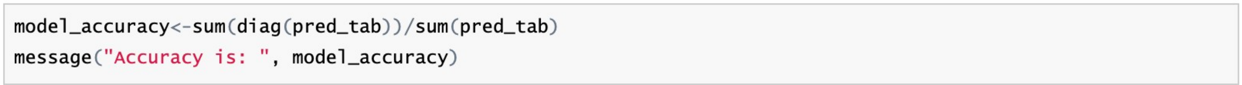
1. The regression coefficients with their values, standard error, Z value and level of significance.
2. AIC and other values, which are used in comparing the model performance



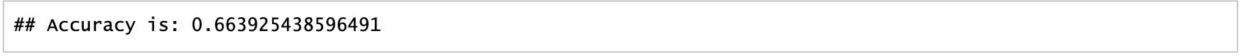
```
pred_tab<-table(pred,cr_data_test$Target)
pred_tab
```



```
##
## pred    1    2    3    4
##    1   32   18    6    4
##    2   44   99   47   51
##    3    0    0    0    0
##    4   80  182  181 1080
```

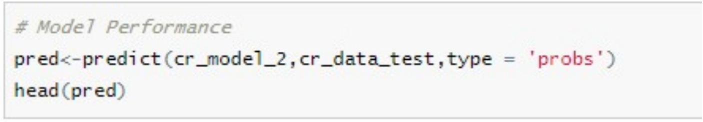


```
model_accuracy<-sum(diag(pred_tab))/sum(pred_tab)
message("Accuracy is: ", model_accuracy)
```

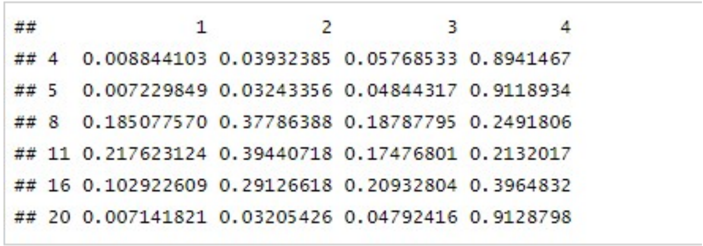


```
## Accuracy is: 0.663925438596491
```

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%.



```
# Model Performance
pred<-predict(cr_model_2,cr_data_test,type = 'probs')
head(pred)
```

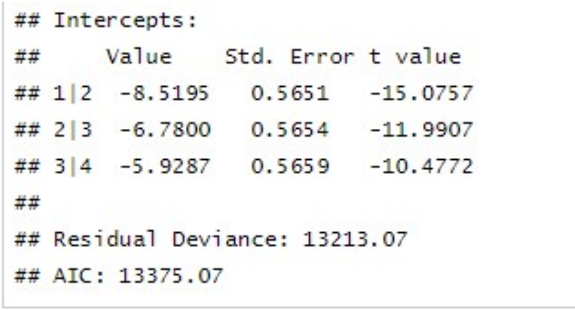


```
##              1          2          3         4
## 4  0.008844103 0.03932385 0.05768533 0.8941467
## 5  0.007229849 0.03243356 0.04844317 0.9118934
## 8  0.185077570 0.37786388 0.18787795 0.2491806
## 11 0.217623124 0.39440718 0.17476801 0.2132017
## 16 0.102922609 0.29126618 0.20932804 0.3964832
## 20 0.007141821 0.03205426 0.04792416 0.9128798
```

The above figure gives us the multinomial probability for categories to all the data points.

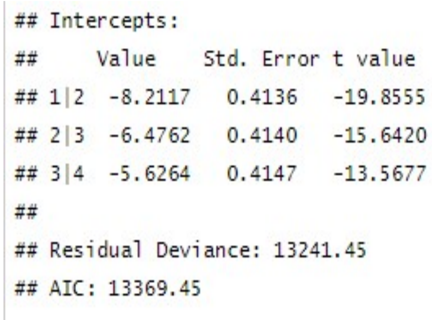
- **MODEL SELECTION**

We determine our model by analyzing the variables and AIC values and select the best model. Model fit analysis examines whether the statistical model employed in an application adequately explains the important features of the data set at hand.



```
## Intercepts:
##     Value    Std. Error t value
## 1|2  -8.5195   0.5651   -15.0757
## 2|3  -6.7800   0.5654   -11.9907
## 3|4  -5.9287   0.5659   -10.4772
##
## Residual Deviance: 13213.07
## AIC: 13375.07
```

Model 1



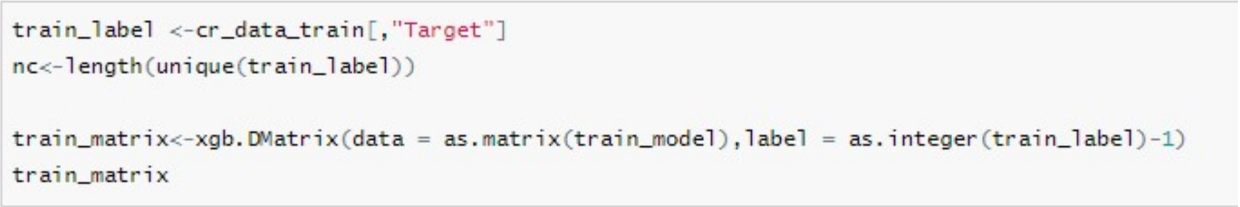
```
## Intercepts:
##     Value    Std. Error t value
## 1|2  -8.2117   0.4136   -19.8555
## 2|3  -6.4762   0.4140   -15.6420
## 3|4  -5.6264   0.4147   -13.5677
##
## Residual Deviance: 13241.45
## AIC: 13369.45
```

Model 2

The model with the lowest AIC value i.e. 13369.45 is the best fit model. From the model selection, we have selected the variables and use the identified variables to build the new model and compare it with the previous model. By comparing the two models we can observe that there is not much improvement in performance of the two models.

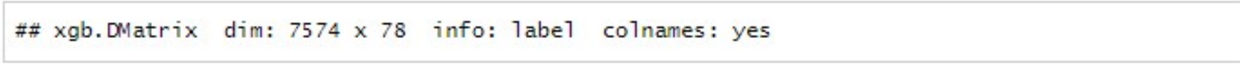
- **MODEL OPTIMIZATION**

Optimization of the model is done using Extreme Gradient Boosting method. 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. Extreme Gradient Boosting is an implementation of gradient boosted decision trees designed for performance and speed.



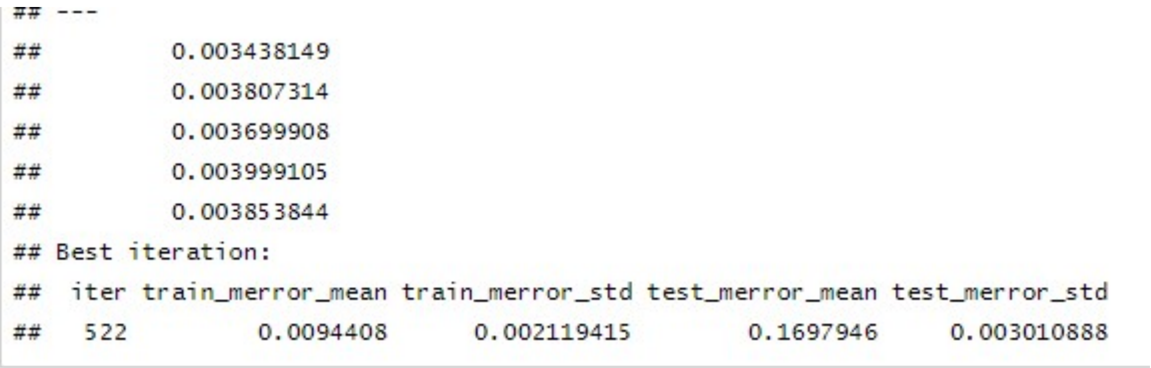
```
train_label <-cr_data_train[,"Target"]
nc<-length(unique(train_label))

train_matrix<-xgb.DMatrix(data = as.matrix(train_model),label = as.integer(train_label)-1)
train_matrix
```

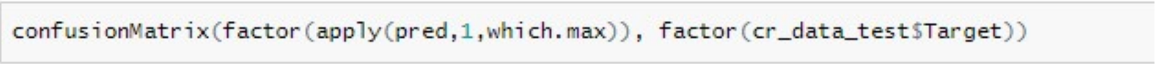


```
## xgb.DMatrix  dim: 7574 x 78  info: label  colnames: yes
```

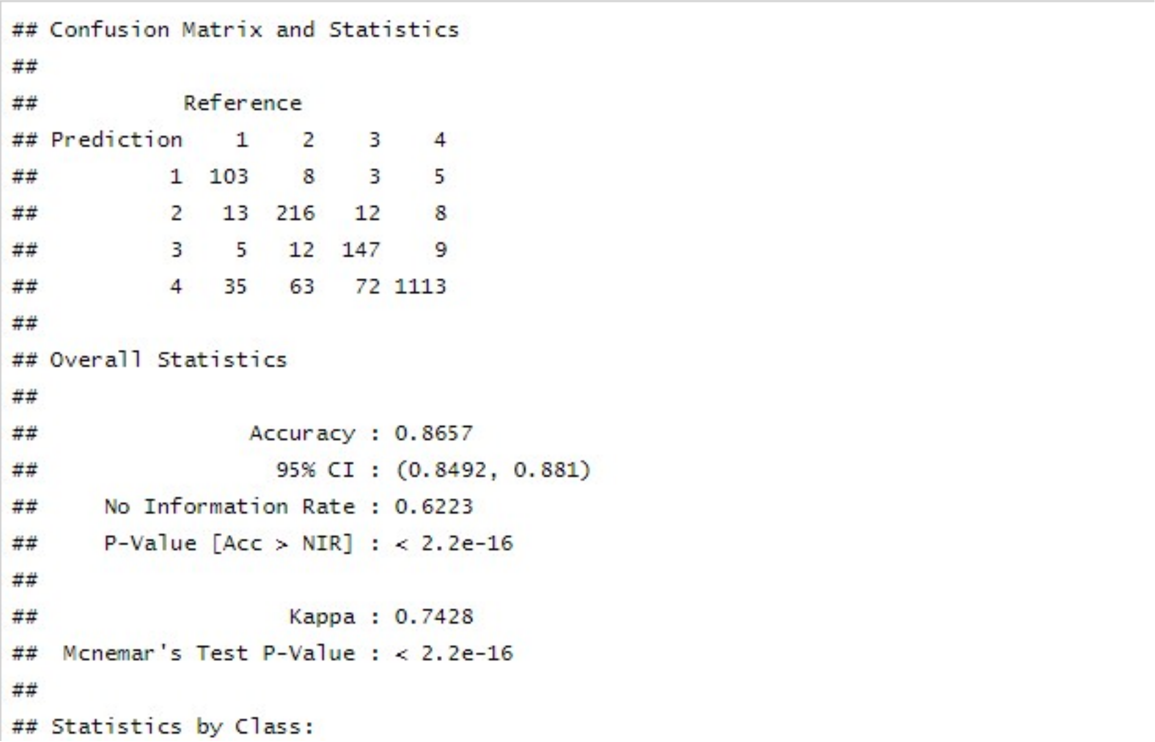
Then we create the parameters for model using list function like number of classes, training rate and created tuning parameters for model to reduce error and overfitting in the model. In our model we have used cross validation method, to find the ideal time to stop the training model and when the validation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



```
## ---
##          0.003438149
##          0.003807314
##          0.003699908
##          0.003999105
##          0.003853844
## Best iteration:
##  iter train_merror_mean train_merror_std test_merror_mean test_merror_std
##   522         0.0094408      0.002119415        0.1697946     0.003010888
```



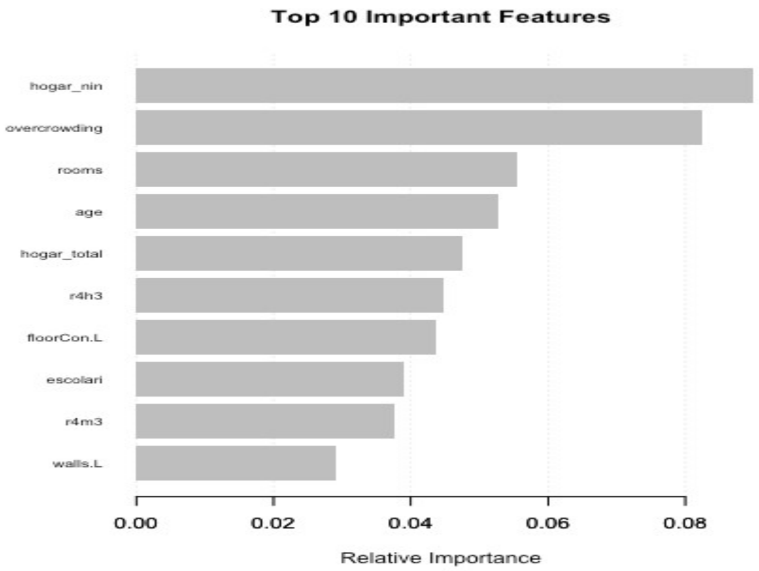
```
confusionMatrix(factor(apply(pred,1,which.max)), factor(cr_data_test$Target))
```



```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    1    2    3    4
##          1  103    8    3    5
##          2   13  216   12    8
##          3    5   12  147    9
##          4   35   63   72 1113
##
## Overall Statistics
##
##                Accuracy : 0.8657
##                  95% CI : (0.8492, 0.881)
##     No Information Rate : 0.6223
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.7428
##  Mcnemar's Test P-Value : < 2.2e-16
##
## Statistics by Class:
```

Using Extreme Gradient Boosting model, we have achieved of 86 % with a 95% confidence interval. The corresponding confusion matrix is also available to determine the accuracy of prediction of each parameter. In the given data the total number of households that belong to category 4 (Non-Vulnerable Household) is more compared to other categories.

Major factors that influence in Predicting the Final Variable:



From the above graph, the most influencing parameter that impacts the poverty level is the number of children between 0 to 19 years old followed by the overcrowding parameter. The more the number of people in an apartment increase there is an increased probability that they will be in the below poverty line. The other factors which influence in predicting the living condition of the household are variables like rooms, age and level of education etc.

CONCLUSION

The model accuracy was found to be 86% and based on factors like sensitivity and specificity, we can propose this model to Inter-American Development Bank to help them analyze socio-economic conditions based on household's characteristics. Government organizations can identify which households need the highest assistance for social welfare using this model. Based on the categorization of poverty level, funds can be allocated to the households that require immediate assistance. The model generated can also be employed to other regions beyond the Costa Rican for assessing the socio-economic conditions.

REFERENCES

Data Analysis & Exploratory Data Analysis (EDA). (n.d.). Retrieved from https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/data-analysis/

Costa Rican Household Poverty Level Prediction. (n.d.). Retrieved from https://www.kaggle.com/c/costa-ricanhousehold-poverty-prediction/data.