Machine Learning Project: Newland

Sara Michetti (m20200626@novaims.unl.pt), Emil Ahmadov (m20201004@novaims.unl.pt),
Nadine Aldesouky (m20200568@novaims.unl.pt), Berfin Sakallioglu (m20200545@novaims.unl.pt),
Ema Mandura (m20200647@novaims.unl.pt)

***Abstract*—The purpose of this project is to help the Newland government impose a taxing system to promote the viability of the planet. This taxing system is applied based on a resident's income compared to all Newland citizens' income. In order to help the government predict which residents fall under which taxing or income group (1 for those who earn above average income and 0 for those who earn below), a team of data scientists performed a rigorous analysis on the existing information about Newland residents. After exploring different machine learning models and algorithms, the team found the Stacking method to be the best predictive model for this study. Through different techniques, this model was optimized and used to perform some hypothesis testing. Finally, it was concluded that older male residents with higher education tend to make up the largest proportion of the Income group 1 which generate higher income and thus will be able to pay higher taxes. The Newland government could use this information to be more selective in its next round of Newland residents while ensuring balance in terms of gender and taxpayers.**

***Index Terms*—machine learning, project, predictive models, stacking, gradient boosting, target encoding**

I. Introduction

The population of Newland is rapidly increasing and as a result, the government will introduce a binary tax rate to ensure the financial viability of this planet. Residents whose income is below the average income of all residents of Newland will pay 15% tax while all the others will pay 30% tax. Thus, the purpose of this study is to identify which residents belong to which taxing group.

The original dataset presents the following variables:

Continuous:

- Income (target variable)
- Birthday
- Years of Education
- Working Hours per week
- Money Received
- Ticket Price

Categorical:

- Native Continent
- Marital Status
- Lives with
- Base Area
- Education Level
- Employment Sector
- Role

In total there are 22400 observations, only 24% of the dataset presents a value of 1 in the target variable. This means that the dataset is imbalanced.

In short, the goal of this project is to build different models that can predict the Income class (0-1) of each observation.

II. Background

For the purpose of this project and the optimization of the final result, different methods and algorithms were tested and implemented. To ensure that the most suitable techniques were used, research was done about algorithms not covered by the subject syllabus.

- `Cramer's V:`

Cramer's V statistic is used to define the association between categorical variables. While its results resemble correlation, the values of Cramer's V range between 0 and 1, with 0 representing no association between variables and 1 representing the complete association. Unlike correlation, association cannot take on negative values. [1]

Given that there are two categorical variables – variable X with values from a1,...,ai and Y with values from b1,...,bi with n observations defined by these variables. With (ai,bj) representing an observation, let nij represent the number of occurrences of this combination in the sample. With row and column totals defined as:

$$n_{i\cdot} = \sum_{j=1}^{J} n_{ij}, \qquad n_{\cdot j} = \sum_{i=1}^{I} n_{ij}.$$

The chi-squared statistic is defined as:

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(n_{ij} - n_{i\cdot} n_{\cdot j}/n)^2}{n_{i\cdot} n_{\cdot j}/n}.$$

Having this denotation, Cramer's V statistic can be computed by a simple transformation:

$$V = \sqrt{\frac{\chi^2/n}{\min(I, J) - 1}}.$$

Cramer's V is symmetrical, and insensitive to swapping X and Y. Indeed, since its sampling distribution is known, standard error and significance can be calculated. [1]

- `Target encoding`

The encoding of categorical variables is a process of transforming one categorical variable into one or more numeric variables to make the data understandable for numeric based

machine learning algorithms. With target encoding, also know as mean encoding, the idea is to use statistics of the target variable for the encoding of the categorical variables.

For mean encoding, for each categorical value, the mean of the target variable is computed. Then, the category is encoded with the calculated mean. This method can be implemented for both binary classification and regression. [2]

One of the downsides of this approach is target leakage which can lead to overfitting. Target leakage occurs when a model is trained based on information that would not be available at the actual time of the prediction. However, there are multiple methods for avoiding this issue. Leaving out the current target value from the target mean reduces target leakage and this method is called leave-one-out encoder. Another option is adding Gaussian noise as a hyperparameter to the encoded value. [2]

Other than target leakage, categories that have only a few examples in the training dataset can also cause issues, as the mean target value could assume extreme values and their encoding might damage the performance of the model. As a solution for this, the category-specific target mean can be mixed with the marginal mean of the whole target variable. [2]

- Local Outlier Factor

In contrast to other outlier detection algorithms, the local outlier factor considers a data point an outlier when it is significantly different from its local neighborhood, rather than from the entire dataset. With this approach, outliers are identified with consideration of the neighborhood density, which is why it performs well in datasets with varying density. [3]

To understand the local outlier factor (LOF), one must consider the concepts of K-distance and K-neighbors, reachability distance, and local reachability density.

With K-distance being the distance between a point and its k-th nearest neighbor, a K-neighborhood is defined as a set of points within a K-distance radius of a point. Further on, reachability distance is considered to be the maximum distance between the K-distance of X1 and the distance between X1 and X2. That means, if point X2 belongs to the K-neighborhood of X1, the K-distance of X1 will be the reachability distance (RD). If not, RD will be the distance between X1 and X2. Local reachability density(LRD) is defined as the inversed average reachability distance of point X to its neighbors. The greater the average reachability distance, the further the point is from its neighbors, and consequently, the lower the density around it. Therefore, the lower the LRD, the further away the nearest cluster is from point X. Finally, local outlier factor (LOF) is the ratio of the average LRD of a K-neighborhood of point X and the LRD of X.

In a regular case, when a point is not an outlier, LOF is close to 1, as the average LRD of a neighborhood and the LRD of a point are almost equal. In the case of an outlier, the average LRD is greater than the point LRD. [3]

Local outlier factor is implemented through sckit learn for Python.

- Grid search

Grid search is a machine learning method used for optimizing the values of hyperparameters. Hyperparameters in a machine learning model are parameters configurated before the model is trained and they control the accuracy of the model. With grid search, an exhaustive search is done over specific model parameters for the right hyperparameters to be chosen. [4]

The python implementation of a grid search is also done through sckit learn.

- Downsampling

One of the biggest data issues that arise with model training is an imbalance. Imbalance refers to skewed class proportions – a significantly large number of data points belong to one class, which is the majority class, while the rest make up the minority class.

This can cause a problem, as with such examples, the training model is mainly focused on the majority class examples and fail to learn enough from the minority class. In the case of imbalanced data, there might be a need for an adapted sampling technique, but only in the case that the model does not perform well on the true distribution of data.

If imbalanced data needs to be handled, downsampling of the majority class can be implemented. To perform downsampling means to take a low subset of the representatives of the majority class. For example, if the original dataset has 1 positive in 100 negatives, downsampling by a factor of 10 would mean taking only 1 in 10 negatives. After this, upweighting should be performed on the data by adding an example weight equal to the downsampling factor to the class. This means that the example that the weight of 10 was assigned to will be treated as 10 times more important during model training.

The benefits of downsampling and upweighting are faster convergence because of more frequent appearances of the minority class, more disk space for the minority class, and calibration that is ensured by upweighting. [5]

III. METHODOLOGY

This section explains the methods used for this research in detail. Procedures for this research were done in the order of data exploration, feature engineering, missing values completion, outlier removal, categorical variables handling, data scaling, synthetic data and finally model building.

- 3.1 Data Exploration

The first thing checked was the data types and how many null values were in the dataset. It was found that there are no null values in the dataset. However, after checking the unique values in each column it was understood that some cells contain the question mark as a value. While other columns had a logical value for the same record, some of the columns contained only question marks. It was decided that these values are null values and question marks were replaced with NaN values to ease further analysis.

The next analysis conducted was checking the distributions of variables. For that purpose, histograms and box plots of only the metric variables were plotted.

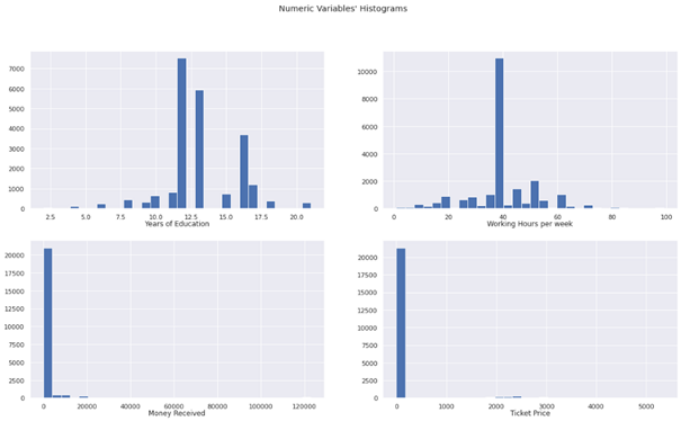


Fig. 1. Numeric Variable's Histograms

One of the inductions made from these distribution plots was that these variables are not normally distributed.

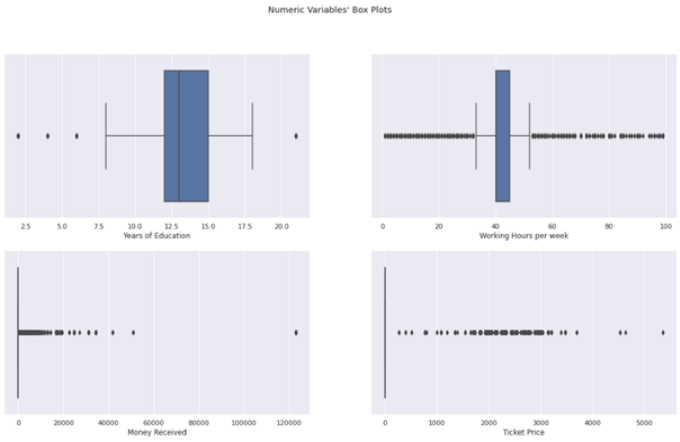


Fig. 2. Numeric Variable's Box Plots

By analyzing the box plots of the metric variables, it can be seen that there are outliers in the data because some points are lying outside of the whiskers. For the categorical variables, frequency and proportion bar charts were plotted to identify which categorical variables have more power of distinguishing. These graphs can be found in appendix A.

- `3.2 Feature Engineering`

In the feature engineering part, the first thing done was to create a Gender variable and identify genders for each record according to their titles. Names column contained "Mr.", "Mrs." and "Miss" titles and this was engineered to find the genders of the records. Gender feature is a binary feature and contains 1 for males, 0 for females.

The next engineered feature was Age. Using the Birthday column from the data, birth year of the records was found and the age of the records was found by substracting birth year from 2048, since 2048 is the year of interest.

The third feature created was the group to which the residents belong. To find out the group of each record, conditions were set as follows: if a person paid a ticket price (group C), if a person receives money (Group B) and if the person did not pay for the ticket and does not get paid (Group A).

- `3.3 Missing Values Completion`

Next, the missing values needed to be filled in. Three columns contained missing values: Role, Employment Sector, and Base Area. Regarding the Role column, K means clustering method was used to determine the most common value among each record's neighbours. When implementing this clustering method, the number of clusters was decided from the inertia plot. Also, Years of Education, Working Hours per Week, and Age variables were used in clustering since they are the variables most related to Role.

For Employment Sector and Base Area variables, mode value was used to fill in the missing values since "Private Sector – Services" category constituted 73% of the Employment Sector column and "Northburry" constituted 91% of the all data for that column.

- `3.4 Outlier Removal`

Since some models are too sensitive to outliers and some are not, two data frames were created; one with outliers and one without outliers. Different methods were used to detect outliers but finally manual detection was most appropriate. The following results were obtained:

- The IQR method suggested dropping 12.9% of the dataset even using a multiplier as high as 11
- The Local Outlier Factor suggested dropping 9.6% of the dataset
- The Isolation Forest method suggested dropping 15% of the dataset

As a result, it was decided to visualize the dataset and use the boxplots of the metric features to detect the outliers. Refer to the plots in 3.1 Data Exploration.

Regarding the Age boxplot, there seemed to be many outliers past the upper bound of the interquartile range so the focus will be on the citizens with an age greater than 80. It was detected that some of these citizens received money from the government to move to Newland. This seems to be a mistake because the government would not pay citizens who are over 80 and working in hard labor industries such as construction for 50 hours per week. The goal is to ensure the viability of Newland and so 80 year old construction workers would contribute to the opposite effect. Nonetheless, to be more critical, 80 year old citizens with high-paid jobs such as Professors were not dropped. Such a record seemed plausible because being a professor is probably an expensive job which is why the government would pay them and would be necessary for the sustainability of Newland. To conclude, 0.27% of the data was dropped as Age outliers since it's a very low proportion and given the objective.

Moreover, the Working Hours per Week contained some outliers too. The determined cut-off region was record that have a value more than 60. This is because 60 hours per week means 12 hours per day all five working days or 8.6 hours per

day all 7 days. This seems to be too extreme and is way over the upper bound of the interquartile range. An assumption was made that the viability of this planet means avoiding illegal exploitation of people and ensuring they can stay healthy, alive, and productive by not being overworked. Thus, given the objective and the small proportion of the dataset (only 3.5%), these records were also dropped.

As for the Years of Education column, there are numerous outliers past the lower bound of the interquartile range. The cut-off range was decided to be citizens with less than 8 years of education as these could be mistakes in the dataset. To confirm this, citizens who did not complete Middle School but somehow got paid by the government to move to Newland were checked. They were mostly old seniors with hard labour jobs while the younger ones were paid more than 50000. This does not seem very logical. Additionally, citizens with less than 8 years of education who paid to move to Newland were checked. This might seem more possible because they might have money from their relatives. However, this small percentage (1.74%) of people with very low education and very low-paying jobs would not be able to pay taxes regardless and would be more of a load on the planet. Thus, these records were dropped so as not to create a wrongful bias in our model.

Finally, following the outlier analysis, 1222 rows were dropped which is equivalent to 5.5% of the dataset. This is to provide higher quality data for the model and analysis.

Database	Original	Outliers	Final
# of observations	22400	1222	21178

- 3.5 Categorical Variables Handling

Some models cannot use categorical variables, so they had to be converted to numerical form. They were transformed using two methods; Target Encoding and One Hot Encoding. In some of the models target encoded variables are used and in some of them one hot encoded variables are used. This is because some models perform better with target encoded variables, such as decision trees. If binary variables are used in decision trees, the tree would grow in one direction which decreases the accuracy of the model [6].Also, compared to one hot encoding, target encoding gives more importance to the categories with more probability of 1.

To check the importance of the categorical variables, Chi-Square analysis for feature selection was applied and Native Continent, Base Area, and Employment Sector proved to not influence the target variable Income. Meaning their impact is insignificant. On the other hand, Lives with, Education Level, Role, and Marital Status are impacting the target variable and need to be used in all the analysis that will follow. To verify the difference between the categories and the target, check the plots in the appendix.

As a result, in order to combine the results of these previous methods, only the variables that repeat in them were used as inputs to the models. Finally, in models two sets of variables were used:

- 1. Continuous variables + one hot encoded categorical (log_features in the notebook)
- 2. Continuous variables + mean target encoded (metrics in the notebook)

- 3.6 Data Scaling

Each of the features in the dataset had a different scale, some with really high values and some with really low values. This would potentially cause problems in models since some models are very sensitive to these ranges. To eliminate this problem, metric features were scaled using MinMax scaler. MinMax scaler was preferred over the StandardScaler since some of the models require only positive values while StandardScaler can produce negative values. Another reason why standard scaler couldn't fit all our continuous variables is that they are not bell-shaped and the scaling would be inaccurate.

Steps 3.2, 3.3, 3.5 and 3.6 are also applied to the Test dataset for consistency.

- 3.7 Synthetic Data

In the dataset, the percentage of ones for the dependent variable is around 24%, which is more than 3 times less than zeros which makes the dataset imbalanced. For this reason, specific methodologies were used to increase the f1 score. One of them was the oversampling method which is called SMOTE (Synthetic Minority Oversampling Technique). This method created synthetic data to decrease the imbalance in the data for better predictions. Also, another method used was downsampling. However, oversampling caused overfitting and downsampling caused underfitting. That is why this synthetic data was finally not used in models.

- 3.8 Model Building

The following section explores the models in detail. This includes the datasets and parameters used for each model as well as the method for choosing the best version of that model. Logistic Regression, Naïve Bayes, K Nearest Neighbors, Decision Trees, Neural Networks, Bagging Classifiers, Gradient Boost and Stacking are the models that were used.

Logistic Regression: Continuous variables and one hot encoded variables were used for this model since these models are designed to work with binary variables. It was expected that basic logistic regression would not be enough considering the complexity of the dataset and to eliminate overfitting. Also, Recursive Feature Elimination (RFE) was used to build a logistic regression model with more important features.

Gaussian Bayes: Gaussian NB works well with normally distributed variables. It was applied on the continuous variables of the dataset, leaving out all the categoricals. This is because statistically, it is incorrect to change binaries or categoricals to Normal distributions. As mentioned earlier, the continuous variables in the dataset are not bell-shaped hence it's not likely that this model will work.

Categorical Bayes: Contrary to the Gaussian Bayes, categorical NB was used only on the categorical variables that were converted with label encoder to make sure that the algorithm could check the distribution for every category in every field. Bayes assumes independence between variables,

which does not help in this analysis. In general, it is difficult to use Bayes with mixed data and ensure it produces good predictions. Continuous variables would need to be discretized to be used with this kind of model and this would lead to loss of information.

KNN: Data without outliers is used in this model since this KNN is sensitive to outliers which would cause lower accuracy. Also, target encoded variables are used since all of the metric variables are scaled in the range of 0 and 1 and binary variables have 0 or 1 values, classifier would give more importance to cluster the points according to binary values. To increase the accuracy of the model, weighted distances method was used. This was also to decrease the sensitivity to noise in the neighborhood. Manhattan distance was not to used as a distance metric in the model since this metric is more suitable for high dimensional data and in the current situation, the data is not of high dimensionality [7]. Instead, the Euclidean distance was used. Also, since the dataset was not too big, it was not very costly to calculate all the distances. For that reason, KNN algorithm was not used with the decision tree algorithm. The best parameter for the number of neighbors was determined using trial and error.

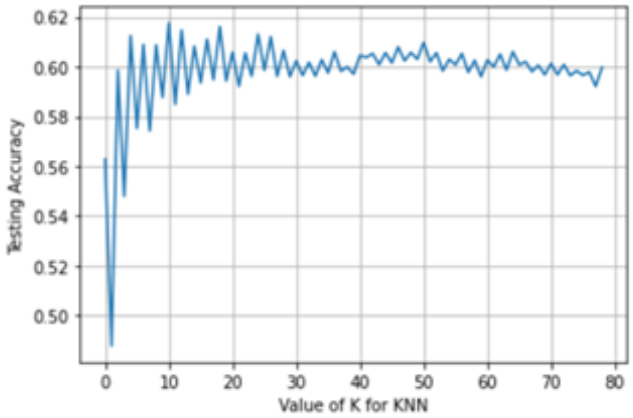


Fig. 3. Testing Accuracy of KNN

Decision Tree: Continuous variables and label encoded variables were used in this model because binary variables would decrease the accuracy of the model. The splitting criteria chosen was 'best', which means that the variable with the highest predictive power would be used for splitting at each node. The criteria for the decision tree was Entropy since there was no major difference when Gini was used. After changing the maximum depth parameter through trial and error, the optimal model was that with a maximum depth of 8 and this model was used to fit the data.

Neural Network: Continuous variables and label encoded variables were used in this model to avoid the high dimensions. The next thing that should be decided was the parameters to find the best model. GridSearch algorithm would be too costly in terms of time if all possible parameters and all possible values for them were fed to the algorithm. To decrease the computation time, some parameters were tested manually to come up with the important parameters which change the score of the model significantly and possible ranges for them.

Bagging: Bagging is the first ensemble learner that was applied to the data. Since the model tries to fit different versions of the base model to different samples from the data, an unstable model must be chosen as the base model. Random Forest Classifier was used for that purpose since it is different for each dataset. Continuous variables and label encoded variables were used in this model since it is a tree based model.

To find the best parameters for the model, the GridSearch algorithm was used. Before running the GridSearch, decisions were made on which parameters to use and ranges for the values of parameters. After checking some parameters manually, ranges for important parameters were decided and fed to GridSearch algorithm. The best parameters from GridSearch were used to build a Random Forest Classifier.

Gradient Boost: This model was the second ensemble learner that was applied to the data. It is also a tree based model and for that reason target encoded variables and continuous variables were used instead of one hot encoded variables.

Important parameters and ranges for important parameters were manually decided and fed to GridSearch algorithm. The best parameters for the model were found to be a learning rate of 0.16, the number of estimators with 170, maximum depth of 3 and maximum number of features with 10.

Stacking: The last model was the Stacking Classifier. In order to decide which base models to use, all of the combinations of previously used models were fitted to the data and Train and Test set scores were calculated for all of them. After getting the scores, the optimal point which avoids overfitting and underfitting was found. First, the models which do not have a major difference between Train and Test set f1 scores were selected. Then, among those, several of the models which have the lowest difference between Train and Test set scores were fitted and the best one was chosen.

IV. RESULTS

In this part, final scores from the models will be provided, the performance of each model will be discussed and the best chosen model will be identified. Later, some hypotheses will be formed about the dataset to distinguish some findings which will then be tested using the best model. Some of the models that are mentioned above are discarded in this section since their results were significantly lower compared to the models below.

To analyze the results of the models, the respective ROC curves are shown below. It can be seen from the graph that the area under the curve is highest for the Gradient Boosting Classifier and the second-best model is Stacking Classifier. There is a very small difference between these two classifiers in terms of AUC. Random Forest Classifier and Neural Network Classifier are third and fourth respectively. Decision Tree classifier performed worst among the models shown, in terms of AUC metric.

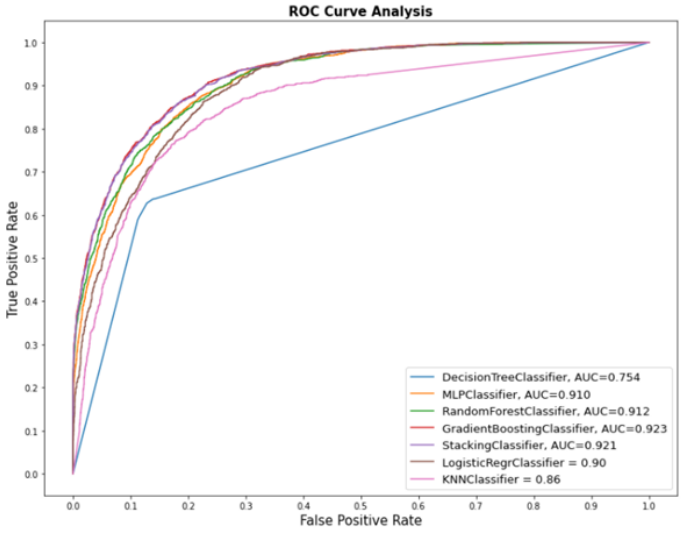


Fig. 4. ROC Curve Analysis

Logistic Regression had a relatively low score. This is because of the non-linear relationship between the Income variable and the explanatory variables. Additionally, there is correlation between the independent variables such as Level of Education and Years of Education. This is a problem because Logistic Regression requires no multicollinearity between the explanatory variables. Given this complex dataset, Logistic Regression generated a low f1 score. [8]

Moreover, the KNN score is not high enough because of the number of neighbors, which in this case was 25. This is quite low, which causes the model to overfit. Since testing accuracy is lower for models that are too complex and/or not complex enough, our model has a lower accuracy compared to the other models. [9]

To decide on the best model, AUC is not enough on its own. Other metrics must be checked such as the f1 score for the Test dataset. These scores are helpful to understand how accurate the models could perform on non-targeted data.

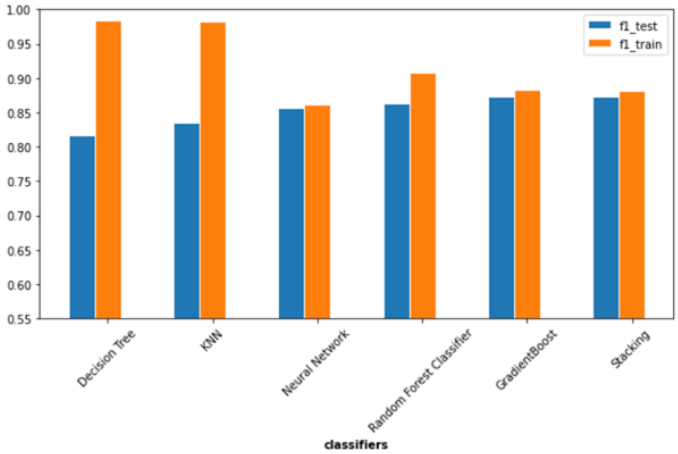


Fig. 5. f1 Scores of the Classifiers

Stacking has the highest score for the Test set with the value of approximately 0.8727, which means on average this classifier is working better than other classifiers on unseen data. Gradient Boosting Classifier has the second-highest score which is 0.8725. The ranking of the other models is the same as the AUC scores one.

The next aspect that needs to be considered is overfitting. To see how models are fitted, scores for train sets are considered. These scores help understand which models are highly dependent on the data which is used to train the models.

Decision Tree and KNN classifiers have an f1 score of approximately 0.98 for the training set. It can be concluded that Decision Tree and KNN classifiers are highly dependent on the train set, which explains why these models performed badly on the non-targeted data. Gradient Boosting algorithm reached 0.8824 as f1 score whereas Stacking classifier has an f1 score of 0.8818, which means that Stacking classifier performs slightly better when it comes to avoiding overfitting. Random Forest Classifier has a score of around 0.90, which means that the model is overfitting. The lowest score is achieved by Neural Network with 0.864.

Regarding the Boosting and Stacking models, the mean of each variable for the two groups of Income were checked. Both models present the same mean and standard deviation for the variables, which does not help in understanding why they differ (tables can be found in the notebook in "GB vs Stacking"). As mentioned earlier, the f1 scores differ, but of its components, just the precision metric slightly changes for the class 0:

	precision	recall	f1-score	support
0	0.90	0.94	0.92	4272
1	0.78	0.64	0.71	1328
accuracy			0.87	5600
macro avg	0.84	0.79	0.81	5600
weighted avg	0.87	0.87	0.87	5600

Fig. 6. Stacking matrix

precision	recall	f1-score	support
0.89	0.94	0.92	4272
0.78	0.64	0.70	1328
		0.87	5600
0.84	0.79	0.81	5600
0.87	0.87	0.87	5600

Fig. 7. Gradient Boosting matrix

Using permutation importance analysis [10] with stacking classifier, the following results were obtained (Fig. 8.).

This shows which features are the ones explaining better the results in the Test set based on the model which was already fitted. The number after the ± measures how performance varied from one-reshuffling to the next. The variables which

Weight	Feature
0.0795 ± 0.0037	Lives_smean_enc
0.0600 ± 0.0039	Role_smean_enc
0.0577 ± 0.0030	Money Received
0.0519 ± 0.0025	Age
0.0293 ± 0.0043	Education_smean_enc
0.0257 ± 0.0010	Working Hours per week
0.0229 ± 0.0047	Years of Education
0.0194 ± 0.0027	Ticket Price
0.0121 ± 0.0020	Employment_smean_enc
0.0043 ± 0.0021	Marital_smean_enc
0.0031 ± 0.0021	Area_smean_enc
0.0012 ± 0.0011	Continent_smean_enc
0.0008 ± 0.0025	Gender

Fig. 8. Permutation importance analysis with stacking classifier

have less impact on the model are Gender, Native Continent, Base area, and Marital Status because they do not change significantly the result on the prediction even when they are shuffled. The variables which have the most impact are Lives With, Role, Money Received, and Age.

Ultimately, Stacking was chosen as the best model because it has the highest f1-score and the most ideal ROC curve.

Indeed, looking at the cleaned dataset, some insights can be made. This is to form some findings or results which can help with the purpose of the study to determine which residents fall under which Income class. These insights will then be tested using the chosen model.

To identify which citizens belong to which taxing group, the most important features in the dataset were visualized.

Looking at the following graphs (Fig. 9.,Fig. 10., Fig. 11.), the following can be concluded:

- 1. Age is not a predictor of whether or not you earn above or below average income.
- 2. The higher the education of the citizen (in terms of years), the more likely they will be to earn above average income.
- 3. Citizens who earn above-average income will mostly be men. We can deduce this because, in the graphs, the female population has a very small percentage of women in the above-average income group, while in the male population a higher percentage of men fall in that group.

The aforementioned hypotheses were formed using the Training dataset, and will next be confirmed or rejected using the Stacking model on the Test set.

Looking at the following results from the Stacking model (Fig. 12.), the first assumption about Age is rejected because it looks like the higher the age, the higher the income. On the other side, the second two assumptions were accepted. This is because Income group 1 has a higher mean for Years of Education and Gender compared to Income group 0. Since males take on value 1 in the Gender column, this higher mean shows that there are more men than women in that income group.

We will confirm this using hypothesis testing below:

Paired Sample T-test for Age:

Since the p-value <0.01, at the 1% significance level, we can reject the null hypothesis which states that there is no

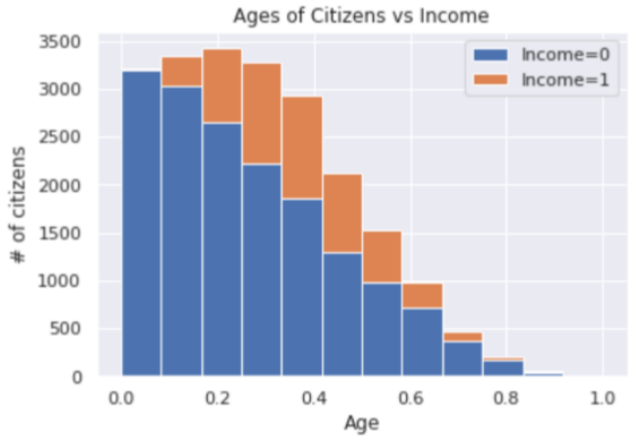


Fig. 9. Ages of Citizens vs. Income

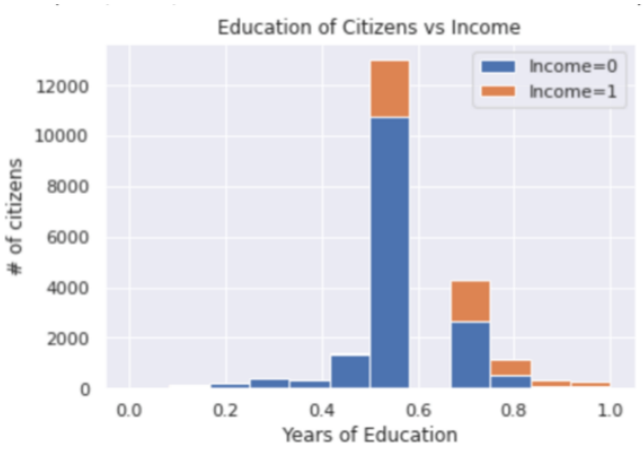


Fig. 10. Education of Citizens vs. Income

Fig. 11. Gender of Citizens vs. Income

y_pred	Years of Education	Age	Gender
0	0.560	0.273	0.636
1	0.701	0.379	0.863

Fig. 12. Stacking Results (Means of Variables)

difference between the mean ages of both income groups. As a result, we can conclude that the difference between the ages of the groups is significant, and the higher the age, the more likely the resident will fall under the Income group 1.

Chi-squared test for Gender:

Since the p-value <0.01, at the 1% significance level, we can reject the null hypothesis which states that the Gender variable is independent of the Income meaning that the proportion of males to females are equal in each category of income. In this case, there is a relationship between the Gender and Income variable. As a result, we can conclude that the proportion of males to females is higher in the income group 1.

Paired Sample T-test for Years of Education:

Since the p-value <0.01, at the 1% significance level, we can reject the null hypothesis which states that there is no difference between the mean Years of Education of both income groups. As a result, we can conclude that the difference between the Years of Education of the groups is significant, and the higher the number of years of education, the more likely the resident will fall under the Income group 1.

V. DISCUSSION

Since there exists some correlation between the independent variables, PCA can be applied to combine some of the categorical and continuous variables together. This would aggregate the impact of those correlated features. To demonstrate, combining the Years of Education and Level of Education variables could be helpful. Additionally, one could focus on the important variables which have high predictive power. For example, an analysis of how the other categorical variables relate to Lives With which is one of the most significant.

The Age variable can also be split into groups to zoom in on which specific age group is most likely to fall under Income group 1. This would make sense because as you grow older you generate more income but only up until a certain age where instead you retire and don't generate any.

Additional data can be collected about the citizens' family loading. In other words, whether or not they have children. This would affect their ability to pay taxes and so their classification in the Income group.

Finally, to balance out the dataset and lose the need for upsampling, additional data regarding citizens from Income group 1 could be collected. This would improve the overall results and analysis of the study.

VI. CONCLUSION

In conclusion, the government of Newland can now use the following variables to predict which residents belong to which taxing or income group. The variable Age also has a considerable impact as older citizens tend to take more higher-paid roles and therefore generate more income. Additionally, the higher the education, the more likely for a citizen to fall under Income group 1. Likewise, residents who are married ten to generate more income than their counterparts. Indeed, the current population is currently split into 24% above average income and 76% below. It is also more populated with men than women.

The Newland government can use this information to better distribute and choose the residents in the next intake. It should increase those who would probably fall under Income group 1 as these residents are necessary to pay more taxes to support the planet. Moreover, for equality purposes, the planet should have an equal proportion of men and women. For maximal viability of the planet, the distribution of Income groups should be more balanced out as well. Finally, a more fair taxing implementation can be reached if Newland government would dive deeper into the residents' situation. For example, their level of expenses (for necessities such as healthcare, etc.) versus their level of income. They must be able to pay this tax and looking at their income is not necessarily sufficient to reach that conclusion.

REFERENCES

[1] "What is Cramér's V statistic?", Statistical Odds Ends, 2020. [Online]. Available: https://statisticaloddsandends.wordpress.com/2020/02/22/what-is-cramers-v-statistic/. [Accessed: 22- Dec- 2020].

[2] "Target Encoding and Bayesian Target Encoding", Medium, 2020. [Online]. Available: https://towardsdatascience.com/target-encoding-and-bayesian-target-encoding-5c6a6c58ae8c. [Accessed: 22- Dec- 2020].

[3] "Local Outlier Factor (LOF)—Algorithm for outlier identification", Medium, 2020. [Online]. Available: https://towardsdatascience.com/local-outlier-factor-lof-algorithm-for-outlier-identification-8efb887d9843. [Accessed: 22- Dec- 2020].

[4] "What Is Grid Search?", Medium, 2020. [Online]. Available: https://medium.com/fintechexplained/what-is-grid-search-c01fe886ef0a. [Accessed: 22- Dec- 2020].

[5] "Imbalanced Data", Google Developers, 2020. [Online]. Available: https://developers.google.com/machine-learning/data-prep/construct/sampling-splitting/imbalanced-data. [Accessed: 22- Dec- 2020].

[6] "One-Hot Encoding is making your Tree-Based Ensembles worse, here's why?", Medium, 2019 ,https://towardsdatascience.com/one-hot-encoding-is-making-your-tree-based-ensembles-worse-heres-why-d64b282b5769. [Accessed: 18 – Dec - 2020].

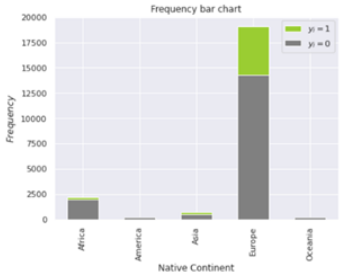
[7] "On the Surprising Behavior of Distance Metrics in High Dimensional Space", Charu C. Aggarwal1, Alexander Hinneburg2, and Daniel A. Keim2, https://bib.dbvis.de/uploadedFiles/155.pdf

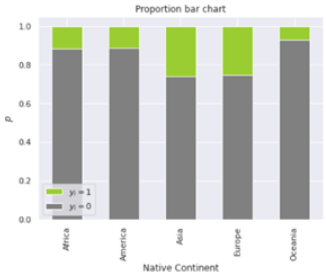
[8] "Advantages and disadvantages of logistic regression", Geeksforgeeks, Amiya Ranjan Rout, 2020, Available: https://www.geeksforgeeks.org/advantages-and-disadvantages-of-logistic-regression/ [Accessed: 20 – Dec – 2020].

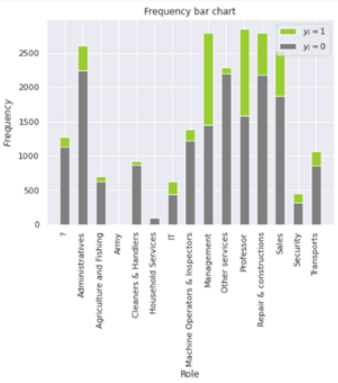
[9] "K-nearest Neighbors (KNN) Classification Model", Ritchieng, Ritchie Ng, unknown, Available: https://www.ritchieng.com/machine-learning-k-nearest-neighbors-knn/ [Accessed: 20 – Dec – 2020].

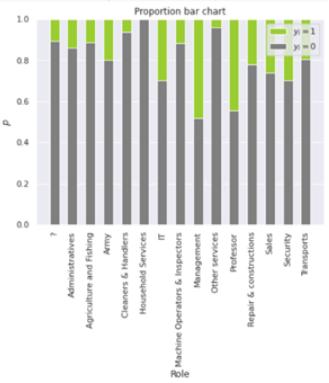
[10] "Permutation Importance", Kaggle , Dan Becket, unknown year [Online Course]. Available: https://www.kaggle.com/dansbecker/permutation-importance [Accessed: 20 – Dec – 2020].

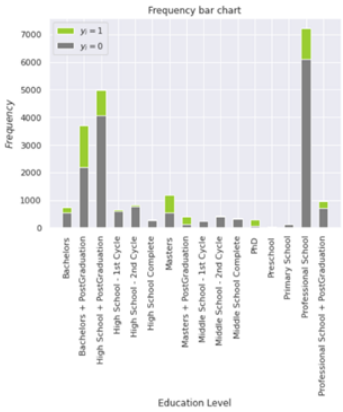
APPENDIX A:

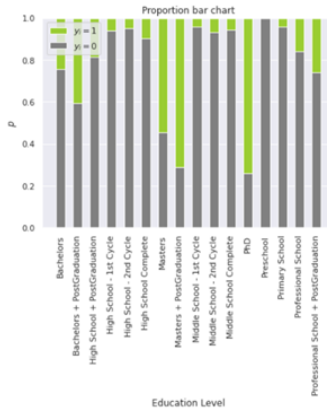

Frequency bar chart
Frequency
Native Continent
Africa
America
Asia
Europe
Oceania
$y_i = 1$
$y_i = 0$

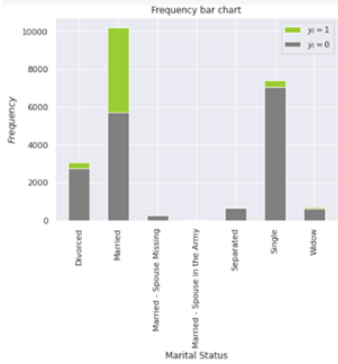

Proportion bar chart
P
Native Continent
Africa
America
Asia
Europe
Oceania
$y_i = 1$
$y_i = 0$

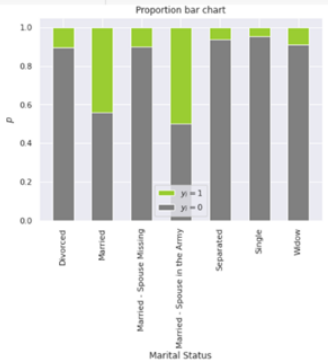

Frequency bar chart
Frequency
Role
?
Administratives
Agriculture and Fishing
Army
Cleaners & Handlers
Household Services
IT
Machine Operators & Inspectors
Management
Other services
Professor
Repair & constructions
Sales
Security
Transports
$y_i = 1$
$y_i = 0$

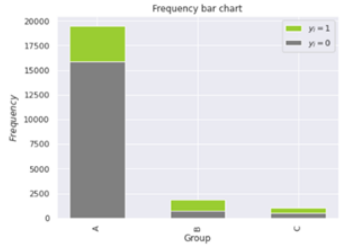

Proportion bar chart
P
Role
?
Administratives
Agriculture and Fishing
Army
Cleaners & Handlers
Household Services
IT
Machine Operators & Inspectors
Management
Other services
Professor
Repair & constructions
Sales
Security
Transports
$y_i = 1$
$y_i = 0$

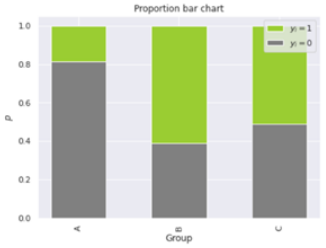

Frequency bar chart
Frequency
Education Level
Bachelors
Bachelors + PostGraduation
High School + PostGraduation
High School - 1st Cycle
High School - 2nd Cycle
High School Complete
Masters
Masters + PostGraduation
Middle School - 1st Cycle
Middle School - 2nd Cycle
Middle School Complete
PhD
Preschool
Primary School
Professional School
Professional School + PostGraduation
$y_i = 1$
$y_i = 0$

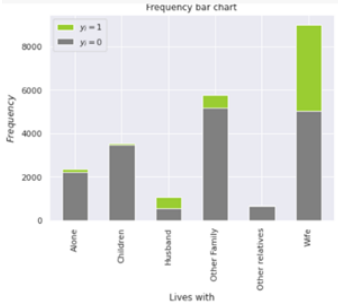

Proportion bar chart
P
Education Level
Bachelors
Bachelors + PostGraduation
High School + PostGraduation
High School - 1st Cycle
High School - 2nd Cycle
High School Complete
Masters
Masters + PostGraduation
Middle School - 1st Cycle
Middle School - 2nd Cycle
Middle School Complete
PhD
Preschool
Primary School
Professional School
Professional School + PostGraduation
$y_i = 1$
$y_i = 0$

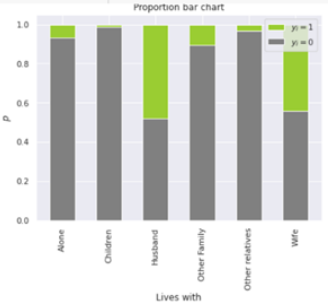

Frequency bar chart
Frequency
Marital Status
Divorced
Married
Married - Spouse Missing
Married - Spouse in the Army
Separated
Single
Widow
$y_i = 1$
$y_i = 0$

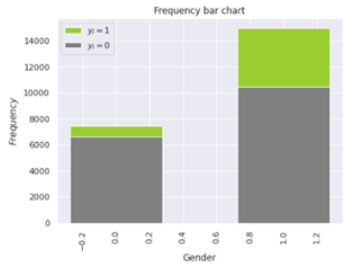

Proportion bar chart
P
Marital Status
Divorced
Married
Married - Spouse Missing
Married - Spouse in the Army
Separated
Single
Widow
$y_i = 1$
$y_i = 0$

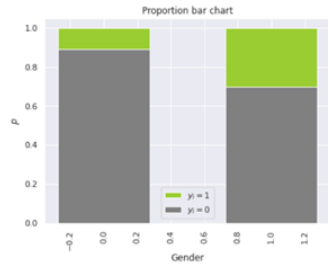

Frequency bar chart
Frequency
Group
A
B
C
$y_i = 1$
$y_i = 0$

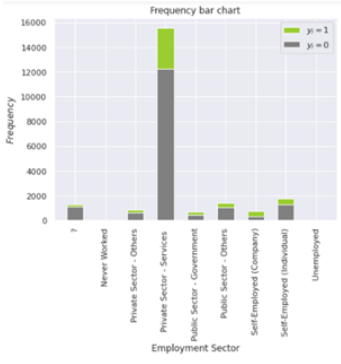

Proportion bar chart
P
Group
A
B
C
$y_i = 1$
$y_i = 0$

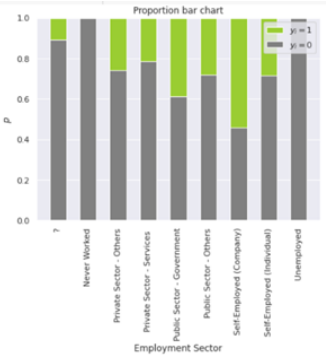

Frequency bar chart
Frequency
Lives with
Alone
Children
Husband
Other Family
Other relatives
Wife
$y_i = 1$
$y_i = 0$

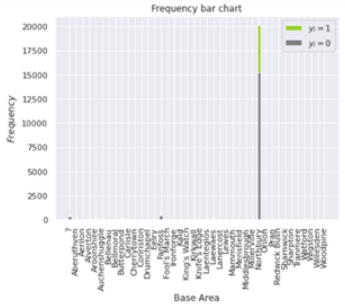

Proportion bar chart
P
Lives with
Alone
Children
Husband
Other Family
Other relatives
Wife
$y_i = 1$
$y_i = 0$

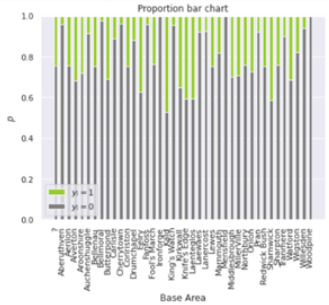

Frequency bar chart
Frequency
Gender
$y_i = 1$
$y_i = 0$


Proportion bar chart
P
Gender
$y_i = 1$
$y_i = 0$


Frequency bar chart
Frequency
Employment Sector
?
Never Worked
Private Sector - Others
Private Sector - Services
Public Sector - Government
Public Sector - Others
Self-Employed (Company)
Self-Employed (Individual)
Unemployed
$y_i = 1$
$y_i = 0$


Proportion bar chart
P
Employment Sector
?
Never Worked
Private Sector - Others
Private Sector - Services
Public Sector - Government
Public Sector - Others
Self-Employed (Company)
Self-Employed (Individual)
Unemployed
$y_i = 1$
$y_i = 0$


Frequency bar chart
Frequency
Base Area
$y_i = 1$
$y_i = 0$


Proportion bar chart
P
Base Area
$y_i = 1$
$y_i = 0$