Final Assignment

FEM11152 – Seminar Data Science for Marketing Analytics

Alexandros Stavropoulos - 643414

2023-01-01

Introduction

In this project I was challenged to investigate how a Finance institute could automate it's Home Loan eligibility process with the use of Machine Learning Algorithms in order to predict in real time if a loan applicant is eligible for a loan or not based on customer's details and information provided while filling an online application form. In order for the customers to be classified based on their loan eligibility, three Machine Learning Algorithms will be used, and evaluated, a K-Nearest Neighbor(KNN) Algorithm,a Random Forest Algorithm and last but not least an Extreme Gradient Boosting (XGBoost) Algorithm in order to determine witch Algorithm can predict better the loan eligibility of customers.

Data Description & Preparations

For this assignment the data set that was used, contains observations of 614 customers and 12 variables that describe their characteristics (Devzohaib 2022). The 12 variable of our data set can provide information about customers gender,marital status,education level,employment and income status, amount of loan, loan terms, credit history, area of the property,if there is a co-applicant and his/hers income level and of course if they are eligible for loan or not. After an initial exploratory analysis of the data set, it was clear that the majority of customers (around 70%) are eligible for loan Figure 1. Due to the fact that,we want our models to perform good in predicting both classes of customers, a re-sampling technique was introduced in our data set. An oversampling technique was used, in order to increase the sample size of non eligible population. The final data set that was used contains 658 observations. Furthermore, observations with missing values were omitted, all numeric variables were normalized and categorical variables were factorized. In order to construct and evaluate our models, the data set was divided into train and test set with 70/30 ration in order to perform an out of sample prediction on test set. The train data set contains 439 observations and the test set contains 219 observations .

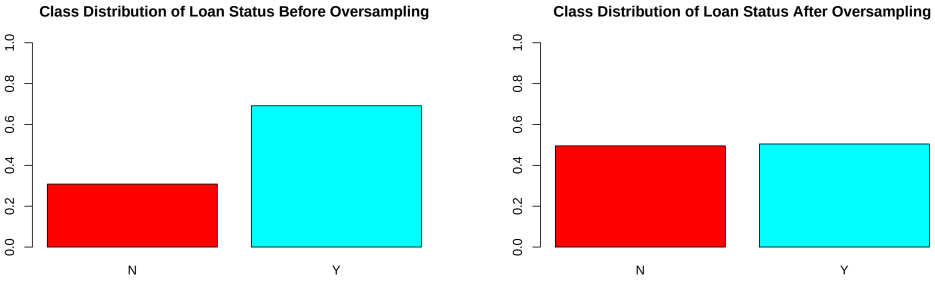

Figure 1: Target Class Distribution

Methods

K-Nearest Neighbors Algorithm The K-Nearest Neighbors Algorithm (KNN) is a non parametric and supervised machine learning algorithm that utilizes the distance and the proximity of a single data object from other similar objects in order to classify it. KNN relies on the assumption that data points with similar features are most probably in the same data area and close to each other. KNN can be used for predicting class labels or for predicting a continuous value. However, the most common use of a KNN model is for classification problems(IBM 2022c).

In a classification setting, KNN uses the term "popularity voting" in order to assign a class label to an individual data point. In simple words it means that based on class with the higher frequency of the data points in proximity with the data point of interest, the class label is assigned ("02-Knn___notes.pdf - Stat 479: Machine Learning Lecture Notes Sebastian Raschka Department of Statistics University of: Course Hero," n.d.). . In many occasion "popularity voting" is confused with "majority voting". In a classification setting of only two labels "popularity" and "majority" means the same thing, but in a more than two labels classification problem a frequency of 50% and more is not necessarily required in order for a class label to be assigned("02-Knn___notes.pdf - Stat 479: Machine Learning Lecture Notes Sebastian Raschka Department of Statistics University of: Course Hero," n.d.).

The decision boundaries of KNN are being established by measuring the distance in data space of points with similar features. The most common distance that is being used, is the Euclidean distance that measures a straight line between data points and can be calculated as $d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$.

The selected number of k nearest neighbors plays a significant role in the model as it defines how many points in proximity will be checked in the classification process. A too small or too high number of k can make predictions unstable and unreliable (IBM 2022c). A common tactic for choosing the number of k is to use it as a hyper parameter and choose the optimal number via cross validation which will be explained later on.

Random Forest The random forest algorithm is an unsupervised machine learning method that belongs to the category of ensemble methods. Ensemble methods creates a strong predicting model that derives from the combination of several weaker methods. Random forest uses the bagging technique and it creates several decision trees in parallel and the final classification is based on the aggregation of the results of every decision tree. Each tree created by random forest is fitted on a subset of the original training data with random replacement (bootstrap) and a small part of this sample is set aside for cross validation later on. The sample that is being set aside is called out of bag sample. The consequence of this procedure is that the final model is trained on a huge diversity of data points with different features. The randomness of bootstrapping also creates uncorrelated decision trees with reduced bias.Random forest can be used for both classification and regression problems. (IBM 2022b).

In random forest there are three parameters that must be selected and play significant role in the final model. The first parameter is the node size that specifies the minimum number of data points that can be included in a node or leaf of a decision tree. A higher number of node size can cause trees to grow too small and a small number of node size can cause trees to grow too big. The second parameter is the number of trees that will be generated and also the number of features sampled. All of these can be used as hyper parameters and cross validated in order to get the optimal number.(L. 1996)

Extreme Gradient Boosting Extreme gradient boosting (XGBoost) is also an ensemble method but unlike with random forest it uses the boosting technique in order to build a strong predicting model. XGBoost in contrast to other boosting algorithms generates trees in parallel,however it follows the same principles in making the final decision due to the fact that every tree learns from the mistakes of previous tree (NVIDIA 2022) . Every tree generated in the ensemble method uses the error residuals of the previous decision tree in order to minimize the overall error of the model. XGBoost utilizes the theory of gradient descent of error. In simple words the model will keep adding new iterations of weak models until the error of the model reaches the minimum possible error level (IBM 2022a) . XGBoost algorithm can be used for both classification and regression problems.

In XGBoost there are many parameters that can be tuned and affect the overall performance of the model. The most notable are eta, gamma and max depth. Eta refers to the learning rate of the model and can be used to avoid over fitting of the model(NVIDIA 2022). Gamma parameter indicates the minimum number of the error reduction of the model in order to make a split or not in a decision tree.Finally max depth determines the size of every tree generated. All of them are can be used as hyper parameters and cross validated.

Cross Validation Cross Validation is the most frequently used statistical method , in order to evaluate the performance of machine learning algorithms and also for tuning the hyper parameters of the models.The most widely used type of cross validation and the one that was used in this project is the K fold cross validation where the training data set is divided into a specified K number of folds or subsets.The model is build on K-1 folds and then evaluated on the remaining fold.The above process continuous until every subset was used as a evaluation set. The average accuracy of all iterations servers as the cross validation accuracy of the model (Strobl C. 2009)

Models Choice Rational The three machine learning methods that were analysed in this section were selected among many others machine learning and deep learning algorithms due to the fact that all of them are strong performers in classification problems and the intantion was to compare the most commonly used “simple” classification algorithm (KNN) against the two most commonly used ensemble methods in a classification setting,Random forest and Extreme gradient boosting.

Analysis & Results

The first model that was fitted on the training set was a KNN model. The number of k for the model was used as a hyper parameter and tuned with 10 fold cross-validation. The final model shown a very high in sample accuracy of 85% Figure 2. In an out of sample prediction the model performed slightly worse than in training set but still with really high performance with accuracy of 81.73% Figure 3.

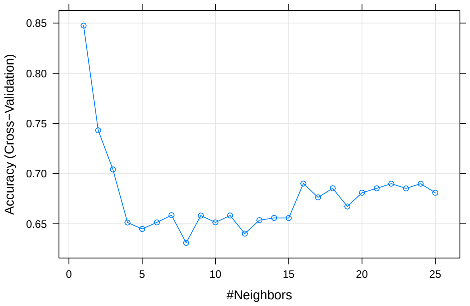

Figure 2: KNN model CV Accuracy

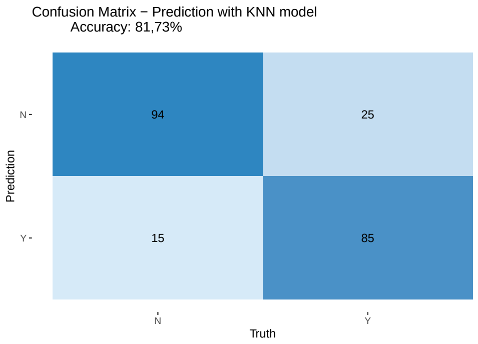

Figure 3: KNN model CV Accuracy

The random forest model was constructed with a node size of 14 observations, 1000 trees while the number of randomly drawn candidate variables $mtry$ was used as hyper-parameter and and tuned with 10 fold cross -validation. The final model performed with an accuracy of 81.05% with $mtry = 10$ on training data Figure 4. After testing the random forest model on test data the performance still good but lower than the KNN model with accuracy of 79.90% Figure 5.

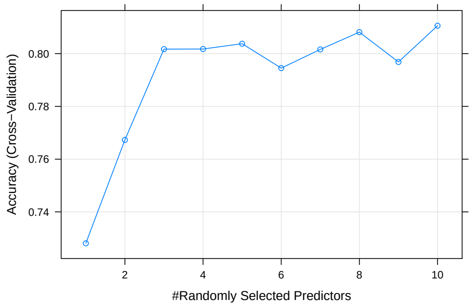

Figure 4: RF model CV Accuracy

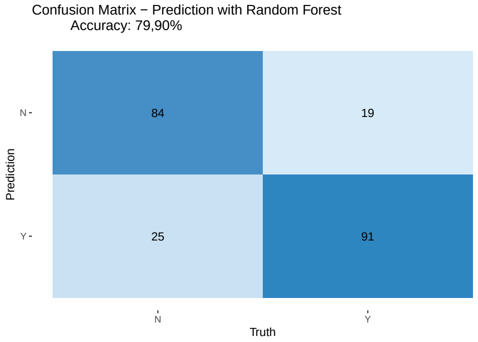

Figure 5: Confusion Matrix Random Forest model

Finally the the XGBoost model was constructed with it's parameters being tuned with 10 fold cross - validation. The final model performed with an accuracy of 87.68% with $eta = 0.4$ and $max_depth = 3$ on training data Figure 6.

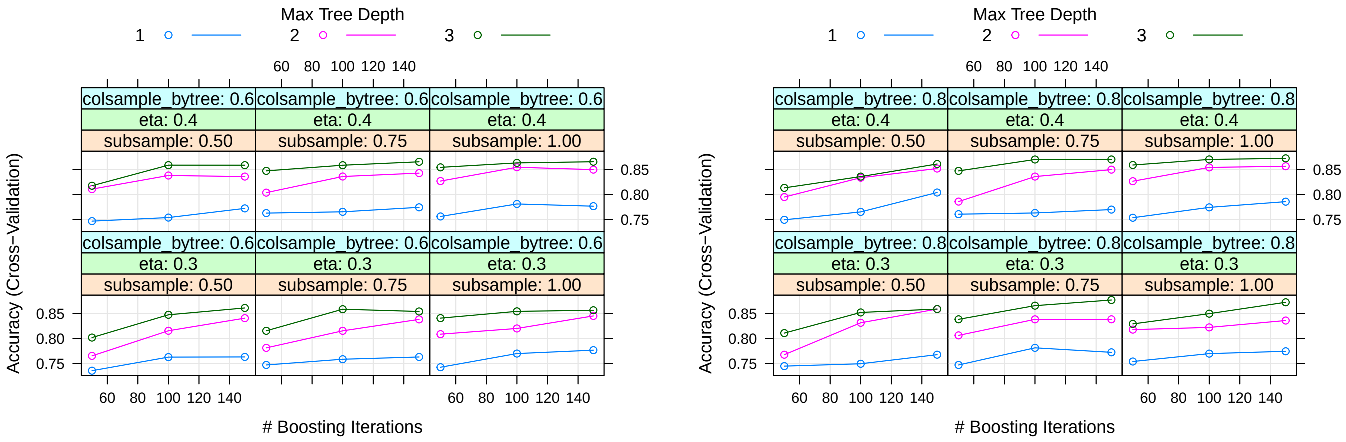

Figure 6: XGBoost model CV Accuracy

After testing the model on test data the performance was very high with accuracy of 83.56% Figure 7

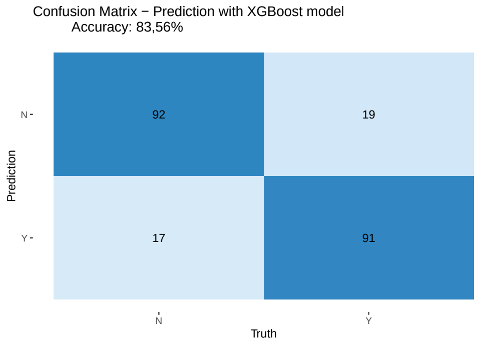

Figure 7: Confusion Matrix XGBoost model

Models Comparison From Table 1, it is clear that while all of the three models perform really good in classifying the loan eligibility of customers, the XGBoost models is a stronger performing model and thus it would be more suitable for automating the loan eligibility process with lower miss classification cost for the company. XGBoost performs really good in predicting both target classes the positive (applicant is eligible for loan) and the negative. Furthermore, it has the lowest overall error rate of the three models. The best model and the one that the company can utilize for their loan eligibility procedure is the XGBoost model.

Table 1: Models Comparison

Models	True Pos	True Neg	False Pos	False Neg	Overall Error Rate %	Accuracy %	Sensitivity %	Specificity %
KNN	85	94	15	25	18.265	81.735	77.273	86.239
RandomForest	91	84	25	19	20.091	79.909	82.727	77.064
XGBoost	91	92	17	19	16.438	83.562	82.727	84.404

Features Importance In order to understand how the best model makes a classification of a customer it is important to know what features of every customer plays the most important role in the classification.

Figure 8 shows that the characteristics of the loan applicant can be divided into two cluster set of features. The first cluster of features consists of the income of the applicant, the amount of money that the applicant request for loan and also his credit history that indicates whether there is another loan at the moment or if he missed any loan payment in the past. The second cluster of features consists of many characteristics with the most notable, if there is a co applicant and his income, the marital status of the applicant and the area of the property. The most important features for classification however, is the income of the applicant which is something very logical, given the fact that an applicant with a higher income is more likely to payoff his depth without a high risk for the institute.

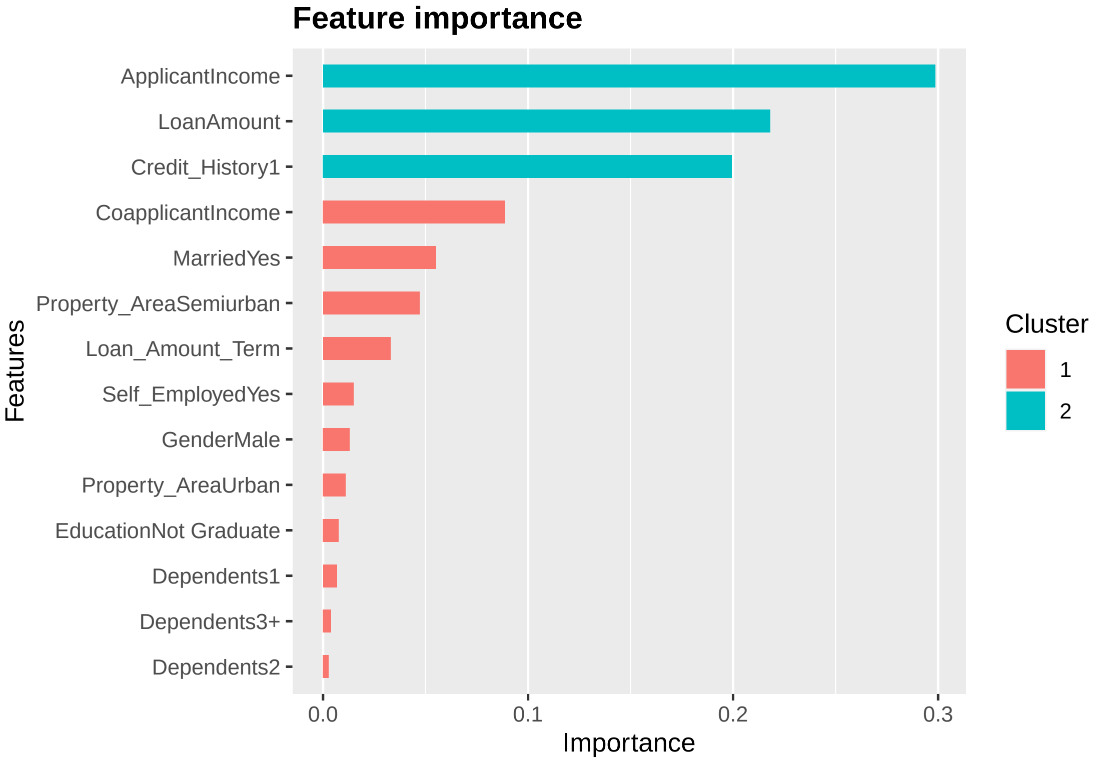

Figure 8: Features impostance of best model

Conclusions

In conclusion, in this project three different machine learning algorithm were examined in order to classify customers for loan eligibility based on customer's details and information provided while filling an online application form and to find out which machine learning algorithm can predict better the loan eligibility in order to help a finance institute to automate it's loan eligibility process. The finding of this project are that customer's details and information provided from the online application form can give the institute all the information they need to predict applicant's loan eligibility. Furthermore all of the models that were examined , can perform with good accuracy however, the XGBoost algorithm has better overall performance over the other models with an overall better error rate. Last but not least, the XGBoost model indicates that the most important feature of an applicant to get approved for a loan is his income level in conjunction with the loan amount that he is applying for.

Bibliography

"02-Knn___notes.pdf - Stat 479: Machine Learning Lecture Notes Sebastian Raschka Department of Statistics University of: Course Hero." n.d. *02-Knn___notes.pdf - STAT 479: Machine Learning Lecture Notes Sebastian Raschka Department of Statistics University of | Course Hero*. https://www.coursehero.com/file/69453649/02-knn-notespdf/.

Devzohaib. 2022. "Eligibility Prediction for Loan." *Kaggle*. https://www.kaggle.com/datasets/devzohaib/eligibility-prediction-for-loan.

IBM. 2022a. "What Is Gradient Descent?" *IBM*. https://www.ibm.com/topics/gradient-descent.

———. 2022b. "What Is Random Forest?" *IBM*. https://www.ibm.com/topics/random-forest.

———. 2022c. "What Is the k-Nearest Neighbors Algorithm?" *IBM*. https://www.ibm.com/topics/knn.

L., BBEIMAN. 1996. https://link.springer.com/content/pdf/10.1007/BF00058655.pdf.

NVIDIA. 2022. "What Is XGBoost?" *NVIDIA Data Science Glossary*. https://www.nvidia.com/en-us/glossary/data-science/xgboost/.

Strobl C., Tutz G., Malley J. 2009. *An Introduction to Recursive Partitioning: Rationale, Application, and Characteristics of Classification and Regression Trees, Bagging, and Random Forests*. Psychological Methods.