Comparing Performance of Machine Learning Algorithms for Default Risk Prediction in Peer to Peer Lending

Yanka Aleksandrova

University of Economics - Varna, Address, Knyaz Boris I blvd 77, Varna, Bulgaria

***Abstract* – The purpose of this research is to evaluate several popular machine learning algorithms for credit scoring for peer to peer lending. The dataset to fit the models is extracted from the official site of Lending Club. Several models have been implemented, including single classifiers (logistic regression, decision tree, multilayer perceptron), homogeneous ensembles (XGBoost, GBM, Random Forest) and heterogeneous ensemble classifiers like Stacked Ensembles. Results show that ensemble classifiers outperform single ones with Stacked Ensemble and XGBoost being the leaders.**

***Keywords* – machine learning, peer to peer lending, credit scoring, ensemble classifiers, XGBoost.**

1. Introduction

The emergence of the sharing economy [1] powered by rapid growth of social networks in all areas [2] has provided the basis for creating communities to share the same values, interests and exchange of ideas and experience. The concept of sharing has also been reflected in the area of peer to peer funding through platforms such as Lending Club, Kickstarter, Bondora, Mintos and more.

DOI: 10.18421/TEM101-16
https://doi.org/10.18421/TEM101-16

Corresponding author: Yanka Aleksandrova,
University of Economics – Varna, Knyaz Boris I blvd 77.
Email: yalexandrova@ue-varna.bg

Received: 08 November 2020.
Revised: 14 January 2021.
Accepted: 20 January 2021.
Published: 27 February 2021.

The expansion in the field of alternative finance in recent years is driven by several prerequisites. First, technology innovations require access to resources provided in a short time frame and at a relatively low cost. Traditional forms of funding cannot provide the necessary funds within these short time frames.

Alternative finances are an appropriate way to overcome the relatively high barrier to obtaining credit through traditional channels like lending by financial institutions, investing through equity, business angels, etc. On the other hand, the ever-decreasing rate of return on deposits in banks and other financial institutions has forced investors to seek alternative forms of investment in their capital resources at higher expected returns.

The global alternative finance market has grown remarkably. In just three years (from 2015 to 2018) its volume has almost doubled, reaching USD 305 billion [3]. Nearly 97% of this sector is formed by peer to peer lending platforms, with peer to peer consumer lending being the most common business model for alternative finances with a share of about 64% of all models in this sector [3].

Peer to peer (P2P) lending brings undeniable advantages for both investors, online platforms, business and individual consumers. However, it also poses substantial risks arising from the specifics of this sector and the dynamic environment. The risk has different dimensions, between which there is a strong dependence, with one of the main threats related to an increase in the share of default loans.

Due to insufficient regulation of the sector and the lack of established rules to reduce credit risk, the responsibility for applying appropriate risk prediction and credit portfolio management methods is placed first and foremost in the hands of online P2P lending platforms. Bad loans are a serious threat to many investors entering the market as well as to online P2P lending platforms and borrowers. This determines the crucial importance of the process of assessing borrowers and predicting the risk of credit default.

At the same time, the advancements in artificial intelligence and machine learning technologies in recent years has led to their ubiquitous application in

all spheres of economy and public life. Successful examples of the use of machine learning to predict and prevent serious threats and unintended consequences are numerous and continuously demonstrate the excellent capabilities of machine learning methods for risk forecasting. In this regard, we formulate the objective of this study as a comparative analysis of the capabilities of different machine learning models for risk prediction in P2P lending platforms. The focus of this research is specifically the risk associated with default loans issued by P2P lending platforms.

Machine Learning models are trained and evaluated using the dataset form the official website of Lending Club – one of the leader platforms for P2P lending.

2. Literature Review

The application of various methods for loan default prediction and credit scoring has been explored in different studies [4], [5]. The main groups of methods applied in this regard are classification models that calculate the probability of a loan applicant stopping the payment or going bankrupt, and survival analysis, which also predicts the time parameters for delaying and suspending payments [6]. Network based models with latent factor models are also applied to credit scoring for potential organizational borrowers of a peer to peer lending platform [7].

The problem of choosing factors to produce a reliable credit score is a subject of many studies [8], [9]. Various factors are used to assess the creditworthiness and probability of default of the loan obligations, such as gender, age, marital status, education, employment length, experience, income [10] income, interest rate, purpose of the loan [11] indebtedness, term of the loan[12], total assets of the borrower [13] customer behavior before and after approval of the loan [14].

Lending Club dataset as a real raw data for machine learning is used in numerous studies. Serrano-Cinca et al. [14] select 18 factor variables classified into five groups: credit grade, credit assessment of the borrower, credit characteristics, loan applicant characteristics, credit history and indebtedness. Employment length at the current position, previous experience with the P2P lending platform, state of address, and FICO is also used [15].

An assessment of the applicability of 41 classification models for assessing the creditworthiness of customers in consumer lending is presented in the survey [5]. The models are divided into three groups:

1. Individual classifiers – neural networks, classification and regression tree, Naïve Bayes classifier, Support Vector Machine (SVM), logistic regression and others.
2. Homogeneous ensemble classifiers – combining several models based on the same algorithm. What these models have in common is that they combine several weak classifiers into one strong.
3. Heterogenous ensemble classifiers - unlike homogeneous ensembles, models based on different algorithms are included.

3. Methodology

The data for this study has been downloaded from the Lending Club site[1]. The data is last updated at the end of July 2020. The original raw dataset structure contains 151 variables. During the data cleaning and understanding phase several transformations on the dataset has been performed, such as:

- Removing of the variables with percentage of missing values greater than 30%. As a result, 58 variables have been dropped.
- Removing of unique variables like url, member_id, id and variables with no or minor variation.
- Converting data to correct format. Several transformations have been performed for dealing with percentages, dates and numeric factors interpreted as text.
- Selecting only variables known at the moment from the credit application. This excludes all factor variables which represents events or conditions happened or occurred after the credit has been funded. The credit assessment performed by Lending Club and expressed in grade, sub_grade, int_rate has not been taken into consideration since the purpose of it is to predict the risk of default with information known before the credit evaluation and approvement form the P2P lending platform.
- Selecting only finished credits. These are credits with two possible states – 0 (Fully Paid) and 1 (Default). Since Lending Club does not provide information whether the credit has been defaulted or paid before the end term, we assume that all finished credits with due date after the second trimester of 2020 are prematurely ended due to some exceptions like default of the borrower or refinancing. That is why we select only finished credits which have reached their final state during the normal term of the credit – 36 or 60 months.

[1] https://www.lendingclub.com/statistics/additional-statistics? [accessed 29 October 2020].

During the feature of engineering phase, we have calculated new variable fico as an average of the low and high margin of the provided FICO score. Employment length has been converted from text to numeric. Another variable has been added to the dataset – dti_loan. It is the ratio of the annual principal payment of the requested loan and annual income of the applicant. It does not consider the interest payment, hence the interest rate, instalment or Lending Club grade.

A new factor variable mths_sinces_first_crl has been added which calculated the number of months from the first opened credit line of the applicant till the issue date of the credit. Another new variable (title_words) has been included as the number of words without stop words the customer has used to describe the purpose of the credit.

For ensuring the quality of the dataset an outlier treatment has been applied. Appropriate methods like outlier removal, capping, discretization have been performed on several variables depending on the outliers percentage and distribution.

Several methods for missing variables imputation have been explored like MICE, Amelia, missForest, kNN. An experiment with 1000 samples revealed that missing values imputations with MICE (Multivariate Imputation by Chained Equations) result in smallest rmse (root mean squared error), mse (mean squared error) and mae (mean absolute error) and that is why MICE has been preferred over the other methods. The cleansed dataset contains 1 467 296 cases of credits issued from 2012 to the end of 2017, and it is described with 28 factor variables (see Table 1). The variables can be divided into five main categories:

Table 1. Description of the factor variables according to Lending Club Data Dictionary

Group	Variable	Description
General profile	emp_length_n*	Employment length (numeric)
	home_ownership	The home ownership status provided by the borrower
Financial profile	annual_inc	The self-reported annual income
	fico*	Average FICO score of the borrower
	mort_acc	Number of mortgage accounts.
	num_bc_tl	Number of bankcard accounts
	num_il_tl	Number of installment accounts
	num_sats	Number of satisfactory accounts
	pub_rec	Number of derogatory public records
	tot_cur_bal	Total current balance of all accounts
	total_bal_ex_mort	Total credit balance excluding mortgage
	total_bc_limit	Total bankcard credit limit
	verification_status	Indicates if income was verified by LC, not verified, or if the income source was verified
Loan characteristics	loan_amnt	The listed amount of the loan applied for by the borrower
	purpose	A category provided by the borrower for the loan request.
	term	The number of payments on the loan - 36 or 60.
	title_words*	Number of words used by the borrower to describe the loan
Indebtedness	bc_util	Total current balance to credit limit for all bankcard accounts.
	dti	Total monthly debt payments excl. mortgage and the requested loan, divided by the monthly income.
	dti_loan*	Principal payment of the requested loan to monthly income
	num_rev_tl_bal_gt_0	Number of revolving trades with balance >0
	revol_bal	Total credit revolving balance
	revol_util	Revolving line utilization rate
Credit history	mnths_since_first_crl	Months since first credit line opened
	mo_sin_old_il_acct	Months since oldest bank installment account opened
	mo_sin_old_rev_tl_op	Months since oldest revolving account opened
	open_acc	The number of open credit lines in the borrower's credit file.
	total_acc	The total number of credit lines in the borrower's credit file

- General profile of the borrower – employment length in years, home ownership;
- Financial profile of the borrower - annual income, FICO score and other financial metrics;
- Credit history of the borrower;
- Loan characteristics – purpose, term, loan amount, title words;
- Indebtedness – debt to income ratios, credit utilization variables and others.

Variables marked with asterisk (*) are calculated and added to the original dataset.

The selected machine learning algorithms for this research are as following:

- Individual classifiers – logistic regression, recursive partitioning and regression tree (rpart), neural networks (multi-layer perceptron with weight decay and deep learning);
- Homogeneous ensembles – eXtreme Gradient Boosting (XGBoost) [16], Gradient Boosting Machine (GBM), Random Forest and Rotation Forest;
- Heterogeneous ensembles – Stacked Ensemble of all models and Stacked Ensemble of Best of Family.

The label variable (loan_status_final) is strongly imbalanced with approximately 80% of the cases being in the positive class (0 = Fully Paid) and 20% - negative cases (1 = Default). Models trained on imbalanced dataset with original distribution of the label variable show exceptional sensitivity (for example 0.9894 for logistic regression), but poor specificity (0.0552). The overall accuracy of the logistic regression model trained on the original dataset is 0.8110, but it is not a reliable metric because of the distribution of the target variable. A more representative metric which should be considered is the balanced accuracy which for logistic regression model is 0.5223, a slightly greater than that of a random classifier.

The problem with the class imbalance has been addressed by applying different balancing techniques. Experiments have been performed to assess different methods such as SMOTE (Synthetic Minority Oversampling Technique), ROSE (Random Over-Sampling Exampling), over sampling of the minority class and under sampling of the majority class. SMOTE and ROSE require substantial computing resources and were not been able to perform on the entire training dataset in a reasonable timeframe of 12 hours on a Microsoft Azure Machine Learning virtual machine with 28 GB RAM. The evaluation metrics of the trained models on smaller random samples using different balancing approaches show however a better performance of traditional under and over sampling techniques and they have been preferred as a way for ensuring equal label distribution of the train dataset. The test dataset is with the original class distribution in order to evaluate the trained models in an environment as close as possible to the real one.

The approach for fitting and evaluation of the machine learning models is shown in Figure 1. All models are trained with 5-fold cross validation and grid search for optimal hyper parameters. The parameter set to be tuned is specific to the chosen algorithm.

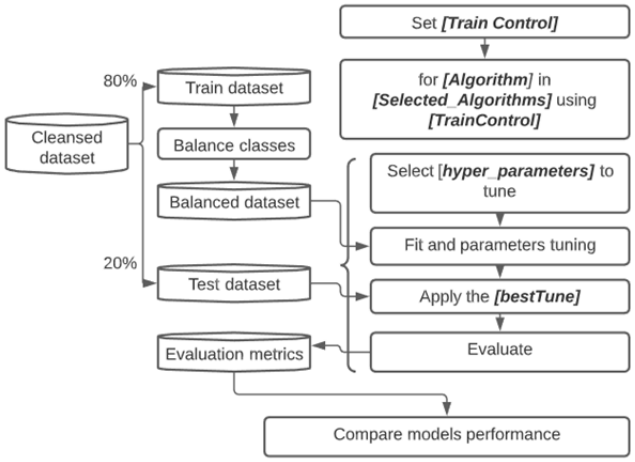

Figure 1. Training and evaluation approach

Depending on the chosen algorithm for every model a certain parameter for regularization has been used to provide robustness and better generalization power. All models are trained and evaluated on same datasets to ensure the comparative performance analysis.

During the evaluation process several important measures derived from the confusion matrix are used like accuracy, balanced accuracy, sensitivity, specificity. For assessing the predictive power of machine learning models we also consider the Kappa coefficient whose calculation adapted to binary classification models is as follows [17]:

$$\text{Kappa} = \frac{total\ Accuracy - random\ Accuracy}{1 - random\ Accuracy}$$

where "random Accuracy" is the expected accuracy of a random guessing model with known base constraints such as the target class distribution. Random Accuracy is thus calculated as:

$$random\ Accuracy = \frac{ActNeg * PredNeg + ActPos * PredPos}{Total\ cases * Total\ cases}$$

where "Act" is an abbreviation of "Actual", "Pred" – "Predicted", "Pos" – "Positive" and "Neg" – "Negative". Kappa coefficient can reveal how better the evaluated model is performing compared to a model which generates predictions by chance.

Another important metric for evaluating classification models is logloss, calculated as follows:

$$logloss = -(y * \log(p) + (1 - y) * \log(1 - p))$$

where y is the value of the positive class (1), p is the probability of a positive prediction given by the classification model. The ideal model has logloss of 0, because it generates always correct predictions with a 100 % probability.

4. Empirical Results

The model training and evaluation has been performed in R using caret and h2o packages. We have used the function train() from the caret package as it offers a flexible and simple interface to train and tune models using more than 230 machine learning algorithms.

4.1. Training and evaluation with caret

The process of hyper parameters tuning depends on the chosen algorithm. The search strategy for the optimal parameter values can be illustrated in Figure 2 where the grid search for optimal values of hyper parameters for XGBoost model is shown. The measure whose values are maximized during the grid search is set to Kappa coefficient. As it is visible in the Figure 2 the maximum value of Kappa = 0.2006 is achieved with max tree depth = 3, eta (learning rate) = 0.3, subsample = 1, colsample_bytree = 0.6.

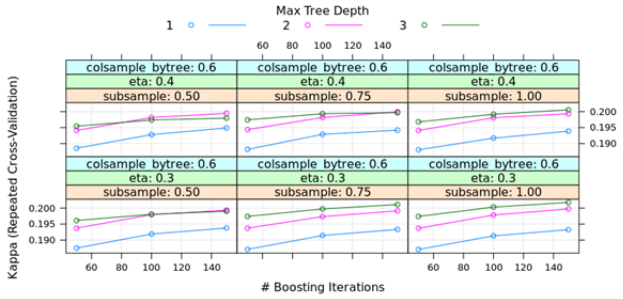

Figure 2. Searching for optimal hyper parameters

We have trained models using six different machine learning algorithms. The best tuned models' performance measures are shown in Table 2. Every model is evaluated with the following measures: accuracy (ACC), Kappa coefficient, sensitivity (Sens), specificity (Spec) and balanced accuracy (BAC). The positive class is 0, i.e. fully paid credit.

All models have accuracy in the range from 0.6465 (Random Forest) to 0.6721 (Rotation Forest). Balanced accuracy is relatively lower varying from 0.6100 (Rpart) to 0.6433 (XGBoost). The model with the highest sensitivity is Rotation Forest which can correctly classify 70.95% from all positive cases.

Specificity is lower than sensitivity for all trained models meaning that models can better distinguish the fully paid credits than default ones. In order to range the models from best to worst we can use Kappa, as an overall measure for model performance. XGBoost has the highest value of Kappa – 0.2006, followed by Multilayer perceptron with weight decay (0.1961) and Random Forest with 0.1900. Taking into consideration all performance measures we can distinguish XGBoost as the best classification model from all included in Table 2.

Table 2. Evaluation of the models tuned with caret::train() function

Model	ACC	Kappa	Sens	Spec	BAC
XGBoost	0.6468	0.2006	0.6490	0.6376	0.6433
GBM	0.6600	0.1864	0.6811	0.5707	0.6259
Random Forest	0.6465	0.1909	0.6540	0.6148	0.6344
Rotation Forest	0.6721	0.1750	0.7095	0.5147	0.6120
Rpart	0.6615	0.1665	0.6930	0.5269	0.6100
MLP	0.6549	0.1961	0.6660	0.6077	0.6368

The feature variable importance in all models reveal more similarities than differences. We have included only the top 5 features according to their relative variable importance in Table 3.

Table 3. Top 5 factor variables ordered by relative variable importance

#	GBM	XGB	Rand. Forest	Rot. Forest	Rpart	MLP
1	term.60 mths	term. 60 mths	dti_loan	term.60 mths	term.60 mths	term.6 0 mths
2	fico	fico	dti	fico	fico	fico
3	dti_loan	dti_lo an	total_bc _limit	purpose .movin g	dti_loan	dti_loa n
4	mort_ac c	loan_ amnt	fico	dti_loan	loan_a mnt	loan_a mnt
5	dti	dti	loan_a mnt	tot_cur _bal	mort_ac c	dti

The most significant variable for all models, except for Random Forest, is credit's term and in particular the 5-year term (60 months). This confirms the observation from the preliminary analysis phase that loans with longer repayment period are significantly riskier. The default rate for 60 months credits is 31% which is double the default rate of 15% for 36 months loans and 60% higher than the overall default rate of 19% for all credits in the dataset.

The Random Forest defines the variables with the strongest relative importance to be those representing the indebtedness of the borrower – dti and dti_loan. These features are identified in the top 5 factors with strongest impact in all six models.

Another important factor influencing the status of the loan is the FICO score, provided by Fair Isaac Corporation. Significant variables among the top five defined by one or more models are also the loan amount, number of mortgage accounts and total limit on all bank cards.

4.2. Training and Evaluation with h2o

We have used the H2O framework for building heterogeneous ensemble machine classifiers. H2O is an open source, in-memory distributed scalable machine learning and analytics platform which allows building machine learning models on big data and facilitates the implementation of those models in production. H2O is accessible to R projects through R interface (h2o-r package).

The concept of stacking ensemble model has been proposed as a weighted combination of many candidate learners to build one strong ensemble learner – the so called "Super Learner" [18], [19]. Stacked Ensemble method in H2O implements the idea of the Super Learner by seeking the optimal combination of different base cross validated machine learning models. During the training of the base models a matrix with dimensions N x L is built where the cross-validated predicted values of all L models on N observations is stored. The matrix is used as "level one" data for training the combination algorithm – a metalearner for the Stacked Ensemble model [20].

When generating predictions on test data Stacked models first collects predictions from the base learners and then feeds those predictions into the metalearner to get the ensemble prediction.

We have used the h2o.automl() function which automates the process of training and tuning the machine learning models including the training of two ensemble heterogeneous models – Stacked Ensemble of all base classifiers and Stacked Ensemble of best of family classifiers. The algorithms used for the base supervised machine learning models are XGBoost, GBM, DRF (Distributed Random Forest), XRT (Extremely Randomized Tree), Deep Learning, GLM (Generalized Linear Model).

During the training an optimal set of classifiers is chosen in which each model is cross validated and tuned with optimal hyper parameters. The best performing models are then stacked into two ensemble models – Stacked Ensemble of All Models and Stacked Ensemble of Best of Family of algorithms.

Twenty-two classification machine learning models have been trained with specified training parameters such as 5-fold cross validation, balance of the target class for dealing with the imbalanced distribution, stopping metric – logloss, stopping tolerance 1e-03, stopping rounds – 3.

The relative importance of the top 10 models in the metalearner of the Stacked Ensemble All Models is shown in Figure 3. Twelve out of all 20 base classifiers have importance in the metalearner greater than 0 and these include 6 XGBoost, 3 Gradient Boosting Machine (GBM), 2 Deep Learning and 1 Distributed Random Forest (DRF) models. Stacked Ensemble metalearner is trained by the default algorithm GLM (Generalized Linear Model).

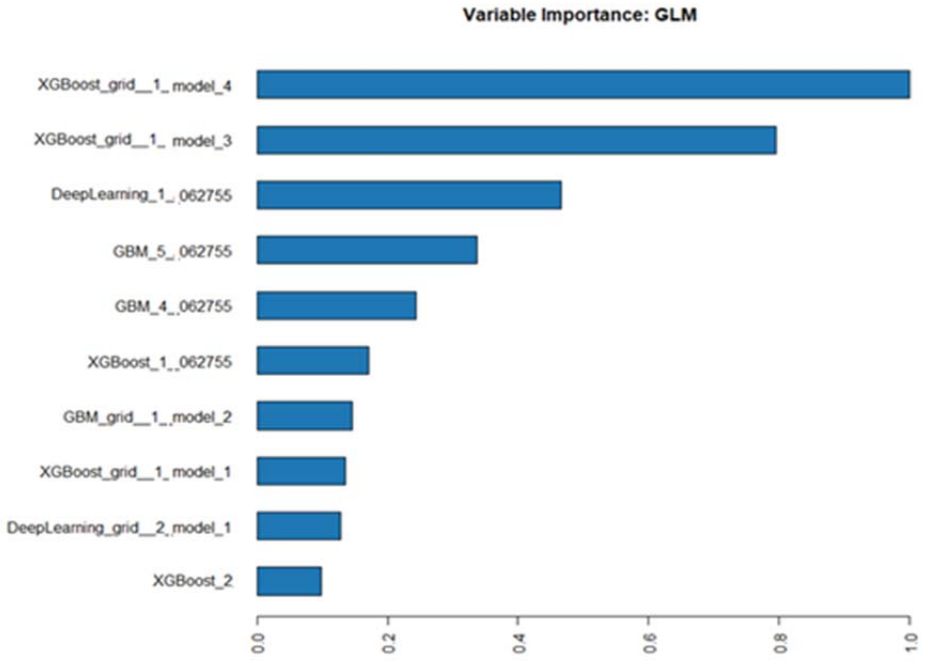

Figure 3. Relative model importance in the metalearner of the Stacked Ensemble All Models

Stacked Ensemble All Models is the leader model of the highest Area Under Curve (AUC), followed by another heterogeneous ensemble classifier – Stacked Ensemble Best of Family – which includes only the best models from each family of algorithms. The performance validation metrics AUC and logloss of the top 10 models are shown in Table 4.

Table 4. Top 10 models ordered by AUC

#	model_id	AUC	logloss
1	StackedEnsemble_AllModels	0.7076	0.4469
2	StackedEnsemble_BestOfFamily	0.7066	0.4474
3	XGBoost_grid__1_model_4	0.7055	0.4453
4	XGBoost_grid__1_model_3	0.7046	0.4456
5	XGBoost_grid__1_model_1	0.7024	0.4465
6	GBM_2	0.7018	0.4468
7	GBM_grid__1_model_1	0.7017	0.4468
8	GBM_1	0.7012	0.4471
9	XGBoost_3	0.7009	0.4471
10	GBM_4	0.7004	0.4475

Variable importance for the best of every family of algorithm is shown in Table 5. The results from Table 5 are also correlated with the ranking of the most significant features of models trained with caret::train function (see Table 3). The feature with the greatest relative importance here is also the term of the credit, identified as such by all models in the Table. The indebtedness characteristics (dti and dti_loan), FICO score, annual income, loan amount, number of months from opening the first revolving bank account and total credit revolving balance complement the list of independent variables that most affect the status of the loan.

Table 5. Top 5 most important features ordered by their relative importance

#	XGB	GBM	DL	GLM	DRF
1	term.36 mths	term	term.60 mths	term.60 mths	term
2	fico	fico	dti_loan	term.36 mths	fico
3	dti_loan	dti_loan	annual_inc	purpose.small business	dti_loan
4	term.60 months	dti	loan_amnt	fico	dti
5	dti	mo_sin_old_rev_tl_op	revol_bal	dti_loan	loan_amnt

Another method for interpreting a machine learning model at a global level except Variable Importance bar charts are Partial Dependency Plots (PDP). They represent the marginal effect that one or two independent variables have on the predictions derived from the training model [21], [22]. These graphs also reveal whether the relationship between factors and result is linear, monotonous or complex. In the case of classification models that display probabilities, PDP can represent a change in the probability of belonging to a target class for different values of the independent variable. We should take however into consideration that these plots assume the independence of the factor variables.

Partial dependency plot on the independent feature term for the top performing models is shown in Figure 4. All models associate the 60 months period with higher risk of default and 36 months loans with significantly lower risk. The mean target response for 36 months term is very close for all models, however for 60 months period models show greater variance of the mean response of the target variable.

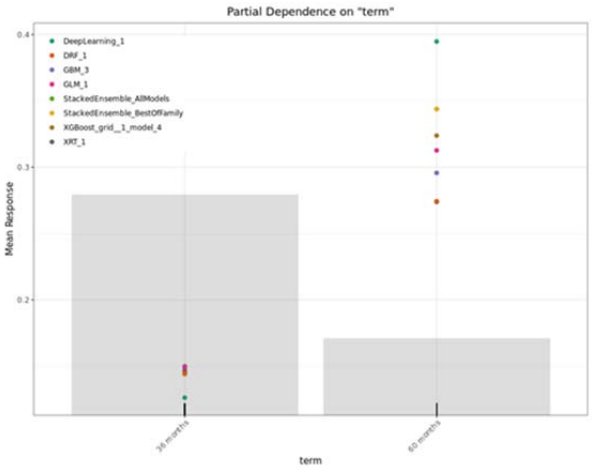

Figure 4. Partial Dependency Plot on term

The marginal effect of the FICO score variable is shown in Figure 5. According to the plot, the default risk decreases with the increase of FICO score value. In models GLM_1 and Deep_Learning_1 the dependency is almost linear, while other models show a non-linear dependency that implies retention and even an increase in risk at FICO values above 800.

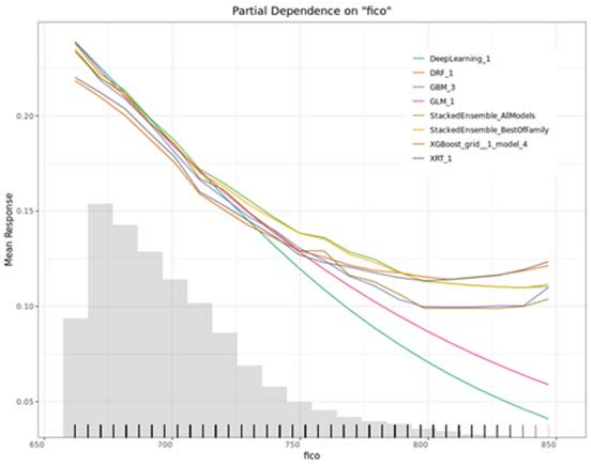

Figure 5. Partial Dependency Plot on FICO

The partial dependency plot for another important feature characterizing the indebtedness of the borrower (dti) is shown in Figure 6. All models associate the high values of this feature with a higher default risk. For the most frequent values in the range of 10% - 25% the marginal effects for all models lie withing close boundaries. For greater values of debt

Table 6. Evaluation metrics for base and stacked ensemble models built in h2o

ID	Model	Acc	Kappa	Sens	Spec	BAC	AUC	logloss
1	StackedEnsemble_AllModels	0.6762	**0.2200**	0.6939	0.6012	**0.6475**	**0.7076**	0.4469
2	Stacked BestOfFamily	0.6875	**0.2196**	0.7059	0.5772	0.6117	**0.7066**	0.4474
3	XGBgrid__1__model_4	0.6875	**0.2231**	0.7138	0.5763	**0.6450**	**0.7055**	**0.4453**
4	XGBgrid__1__model_3	0.6824	**0.2196**	0.7059	0.5830	**0.6444**	0.7046	**0.4456**
5	XGBoost_grid_1_model_1	0.6827	0.2175	0.7077	0.5772	0.6425	0.7024	**0.4465**
6	GBM_2	0.4471	0.1111	0.3451	0.8782	0.6117	0.7018	0.4468
7	GBM_grid_1_model_1	0.4452	0.1106	0.3422	0.8806	0.6114	0.7017	0.4468
8	GBM_1	0.4460	0.1103	0.3437	0.8782	0.6110	0.7012	0.4471
9	XGBoost_3	0.6771	0.2132	0.6992	0.5836	0.6414	0.7009	0.4471
10	GBM_4	0.5049	0.1378	0.4287	0.8270	0.6278	0.7004	0.4475
11	GBM_3	0.4668	0.1196	0.3736	0.8605	0.6171	0.7003	0.4476
12	Deep Learning_1	0.4225	0.0996	0.3105	**0.8960**	0.6032	0.6998	0.4482
13	GBM_5	0.5398	0.1528	0.4813	0.7871	0.6342	0.6980	0.4486
14	XGBoost_1	**0.7023**	0.2185	**0.7443**	0.5248	0.6345	0.6956	0.4501
15	GLM_1	0.6607	0.1998	0.6754	0.5984	0.6369	0.6918	0.4509
16	DeepLearning_gr_1_mod_2	0.4005	0.0870	0.2816	**0.9029**	0.5923	0.6879	0.4547
17	GBM_grid_1_model_1	0.6385	0.1836	0.6430	0.6198	0.6314	0.6848	0.4559
18	XRT_1	0.4878	0.1201	0.4102	0.8158	0.6130	0.6807	0.4568
19	DRF_1	0.4904	0.1208	0.4144	0.8120	0.6132	0.6804	0.4576
20	XGBoost_2	**0.7314**	0.1952	**0.8096**	0.4009	0.6052	0.6727	0.4622
21	DeepLearning_gr_2_mod_5	0.4155	0.0921	0.3037	**0.8879**	0.5958	0.6495	0.4749
22	XGBoost_1	**0.7657**	0.1244	**0.8971**	0.2105	0.5538	0.6351	0.5732

to income ratio however there are big fluctuations in the mean target response between explored models.

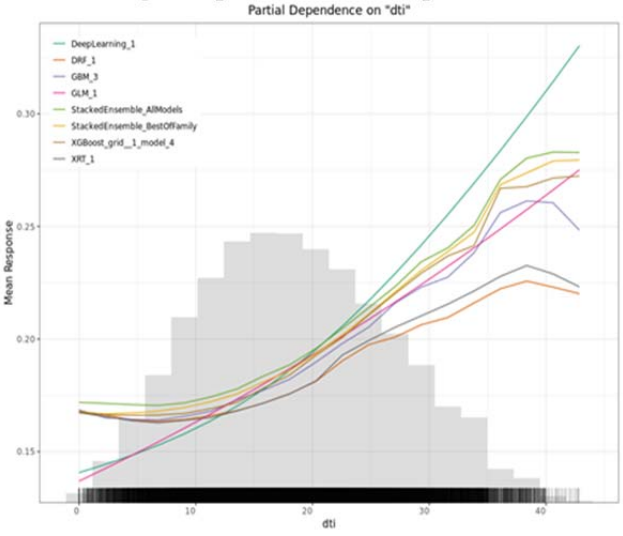

Figure 6. Partial Dependency Plot on debt-to-income

The evaluation of models' performance on validation and train sets is shown in Table 6. The measures AUC (Area Under Curve) and logloss are calculated from validation sets and the rest measures – Acc (accuracy), Kappa, Sens (sensitivity), Spec (specificity), BAC (balanced accuracy) – from the test dataset. For every measure the top 3 values are highlighted – for the logloss the highlighted values are the bottom 3, and for the rest columns – top 3 values.

All models included in Table 6 are trained and tuned by h2o.automl() function. There are different models from each family of algorithm with optimal sets of hyperparameters. The leaderboard shows the best performing models ordered by AUC. According to this evaluation measure the best models are Stacked Ensemble All Models, Stacked Best Of Family and XGBoost model tuned with grid search (model_4).

It is difficult to rank the models as there are several possible measures for ordering. However, we think that according to the results from Table 6 several **conclusions** can be drawn:

- There is no obvious winner – a model which dominates significantly in every measure of its performance. Instead we can divide the models in several groups according to their evaluation metrics – 1) models with good overall performance expressed in relatively high values of Kappa, balanced accuracy and AUC and with low values of logloss. 2) models with excellent sensitivity above 0.70 but poor specificity and 3)

models with exceptional abilities to classify correctly the default loans expressed in specificity above 0.89 but with poor sensitivity to fully paid loans.

- Models trained on one algorithm show different performances when applied to the validation and test set. This can be explained with the crucial role of the optimal values of hyperparameters which can significantly change the generalization and predictive power of the model.
- When using Kappa and AUC as measures for the overall model performance we can conclude that the best performing models are Stacked Ensembles (id 1 and 2) and the top 2 XGBoost models (id 3 and 4). Stacked Ensemble All Models, and the two XGBoost models are ones with the highest values of balanced accuracy.
- The models with the three lowest values of logloss are built on XGBoost algorithm (id 3,4 and 5) which reveals that this algorithm generates predictions with minimum differences between prediction probabilities and actual value of the target variable. On the other hand, the model with the highest logloss is also a XGBoost model (id 22). This model however is with highest Accuracy (0.7657) and Sensitivity (0.8971).
- All the three deep learning models reveal exceptional specificity meaning they can correctly classify more than 89% from default loans. At the same time their sensitivity is very low – they can identify correctly approximately 30% from loans that would be fully paid within the set term.
- GBM, DRF and XRT models show relatively poor performance with low values of Kappa, accuracy not greater than 0.5 and logloss in the middle range of all models.

The models included in Table 6 are all built using the h2o.automl function in h2o -r package as the purpose of this research is to provide a base to compare the performance of different machine learning algorithms. The training and tuning of 22 models using the h2o framework produced models with good predictive power and optimal values of evaluation metrics depending on the training dataset. At the same time, it is worth mentioning that the training process in h2o required less computational resources than training and tuning using the caret package. For comparison the time needed to train and tune one model with caret::train function was approximately 2-3 hours while the whole process of tuning the 22 models in h2o took less than 3 hours. All the experiments have been implemented on the same virtual machine – Microsoft Azure Machine Learning with compute machine STANDARD_DS12_V2 (4 Cores, 28 GB RAM, 56 GB Disk).

5. Conclusion

The empirical results show that machine learning models can be applied successfully to predict default loans in peer to peer lending platforms. To achieve good performance and predictive power we can give the following recommendations:

- Ensure the needed **data quality** by outliers treatment and missing values imputation. It is recommended to test different methods and choose the appropriate one depending on data.
- When the target variable has imbalanced distribution it is necessary to address this issue by **implementing techniques for balancing the target classes**. These techniques however have to be applied only to the training dataset. The test dataset should keep the original class distribution.
- **Model tuning** to find the best values for hyperparameters is mandatory to optimize the model performance.
- During the modeling phase of every machine learning project it is necessary to train and tune models using different algorithms and **compare the models performance**. The evaluation of models should take into consideration different measures like accuracy, balanced accuracy, sensitivity, specificity, Kappa coefficient, AUC, and logloss.

The results from the implemented experiments in this paper show that ensemble models outperform the individual classifiers in terms of generalization on the test set. Several ensemble classifiers have been evaluated and according to their results we can draw the **conclusions** that:

1. Heterogeneous ensemble classifiers and homogeneous ensembles like boosted tree ensembles have better overall performance than individual classifiers and bagged ensembles like Random Forest and Rotation Forest;
2. XGBoost is the leader from all homogeneous ensembles with overall performance competing that of Stacked Ensembles;
3. Despite the noted difference in performance measures there are more similarities than differences in the identification of the most important variables and partial dependency plots between different models. Variable importance and partial dependency plots are methods for global interpretation of the model and should be considered along with the evaluation metrics when choosing the best machine learning model.

Future research on this topic would focus on profit/cost analysis on models' performance. As it was evident from the results, some of the models are more capable of identifying the positive cases while others classify more correctly the negative cases.

Misclassification errors have associated costs which could be direct in case the loan has been issued to a borrower who would not pay back in time or indirect when the P2P platform would not profit from taxes and interest payments.

The growing demand for explainable machine learning models [23], [24] puts the focus on building not only models with good predictive power but providing adequate models and tools for model explanation and interpretation [25]. With this regard the next development of this research would be on applying different agnostic models for interpretation and explanation of predictions and mechanism of the best performing machine learning models.

Acknowledgements

This research has received funding from the European Union's Horizon 2020 research and innovation program FIN-TECH: A Financial supervision and Technology compliance training programme under the grant agreement No 825215 (Topic: ICT-35-2018, Type of action: CSA).

References

[1]. Hamari, J., Sjöklint, M., & Ukkonen, A. (2016). The sharing economy: Why people participate in collaborative consumption. *Journal of the association for information science and technology*, *67*(9), 2047-2059. doi: 10.1002/asi.23552.

[2]. Parusheva, S. (2019). Social Media Banking Usage From Banks' Perspective. *International Journal of E-Business Research (IJEBR)*, *15*(1), 38-54. doi:10.4018/IJEBR.2019010103

[3]. Ziegler, T., & Shneor, R. (2020). *The Global Alternative Finance Market Benchmarking Reporttrends, Opportunities and Challenges for Lending, Equity, and Non-investment Alternative Finance Models*. Cambridge Centre for Alternative Finance. Retrieved from: https://www.jbs.cam.ac.uk/wp-content/uploads/2020/08/2020-04-22-ccaf-global-alternative-finance-market-benchmarking-report.pdf [accessed 29 October 2020]

[4]. Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., & Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the operational research society*, *54*(6), 627-635. doi:10.1057/palgrave.jors.2601545.

[5]. Lessmann, S., Baesens, B., Seow, H. V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, *247*(1), 124-136. doi: 10.1016/j.ejor.2015.05.030.

[6]. Ahelegbey, D. F., Giudici, P., & Hadji-Misheva, B. (2019). Latent factor models for credit scoring in P2P systems. *Physica A: Statistical Mechanics and its Applications*, *522*, 112-121. doi: 10.2139/ssrn.3325231 .

[7]. Abdou, H. A., & Pointon, J. (2011). Credit scoring, statistical techniques and evaluation criteria: a review of the literature. *Intelligent systems in accounting, finance and management*, *18*(2-3), 59-88. doi: 10.1002/isaf.325

[8]. Berger, A. N., & Black, L. K. (2011). Bank size, lending technologies, and small business finance. *Journal of Banking & Finance*, *35*(3), 724-735.

[9]. Šušteršič, M., Mramor, D., & Zupan, J. (2009). Consumer credit scoring models with limited data. *Expert Systems with Applications*, *36*(3), 4736-4744. doi:10.1016/j.eswa.2008.06.016

[10]. Agapitov, A., Lakman, I., Maksimenko, Z., & Efimenko, N. (2019). An approach to developing a scoring system for peer-to-peer (P2p) lending platform. In *Springer Proceedings in Mathematics and Statistics* (pp. 347-357). doi:10.1007/978-3-030-21158-5_26

[11]. Polena, M., & Regner, T. (2016). *Determinants of borrowers' default in P2P lending under consideration of the loan risk class* (No. 2016-023). Friedrich-Schiller-University Jena. Retrieved from: https://www.econstor.eu/handle/10419/148902 [accessed 25 August 2020]

[12]. Zhou, G., Zhang, Y., & Luo, S. (2018). P2P Network Lending, Loss Given Default and Credit Risks. *Sustainability*, *10*(4), 1-15. doi: 10.3390/su10041010.

[13]. Wang, Z., Jiang, C., Ding, Y., Lyu, X., & Liu, Y. (2018). A novel behavioral scoring model for estimating probability of default over time in peer-to-peer lending. *Electronic Commerce Research and Applications*, *27*, 74-82.

[14]. Serrano-Cinca, C., Gutiérrez-Nieto, B., & López-Palacios, L. (2015). Determinants of Default in P2P Lending. *PLOS ONE*, *10*(10), 1-22. doi:10.1371/journal.pone.0139427 .

[15]. Ariza-Garzón, M. J., Arroyo, J., Caparrini, A., & Segovia-Vargas, M. J. (2020). Explainability of a machine learning granting scoring model in peer-to-peer lending. *Ieee Access*, *8*, 64873-64890. doi:10.1109/ACCESS.2020.2984412

[16]. Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794). doi:10.1145/2939672.2939785 .

[17]. Shmueli, B. (2019). *Multi-Class Metrics Made Simple, Part III: the Kappa Score (aka Cohen's Kappa Coefficient)*. Retrieved from: https://towardsdatascience.com/multi-class-metrics-made-simple-the-kappa-score-aka-cohens-kappa-coefficient-bdea137af09c [accessed 29 October 2020]

[18]. Van der Laan, M. J., Polley, E. C., & Hubbard, A. E. (2007). Super learner. *Statistical applications in genetics and molecular biology*, *6*(1). doi:10.2202/1544-6115.1309 .

[19]. Polley, E., & van der Laan, M. (2010). Super Learner in Prediction. *U.C. Berkeley Division of Biostatistics Working Paper Series*.

[20]. H2O.ai. (2020). Stacked Ensembles. Retrieved from: https://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/stacked-ensembles.html [accessed 30 October 2020].

[21]. Apley, D. W., & Zhu, J. (2016). Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models. *arXiv preprint arXiv:1612.08468*.

[22]. Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189-1232.

[23]. Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, *267*, 1-38. https://doi.org/10.1016/j.artint.2018.07.007

[24]. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135-1144).

[25]. Lipton, Z. C. (2018). The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue, 16(3), 31-57.