Loan prediction and risk assessment through ML methodologies

Introduction

The aim of this project consist of analyze a dataset containing economic and demographic information about clients applying for loans. The dataset includes various features related to costumer's financial behavior and a target variable indicating whether the loan application was accepted or denied. Additionally, the dataset provides a risk score for clients, reflecting their creditworthiness. The idea is to apply different supervised Machine Learning methodologies in order to face two main aspects:

- **Classification:** Use machine learning models to predict whether a loan application is accepted or denied based on the provided features.
- **Regression:** Implementing linear regression to estimate the risk score of costumers based on their attributes.

The objective is to train and test various types of models, measure their performance through a rigorous process of model selection and validation, and ultimately identify the best-performing model. Once the optimal model has been identified, it will be tested on unseen data, without further training phase, to evaluate its ability to generalize and assess whether its performance holds steady or if signs of overfitting emerge. To gain deeper insights from the data and the relationships between the variables, the project also involves:

- **Correlation Analysis:** Examining the relationships between features and target variables to identify the most influential factors affecting loan approval.
- **Data Distribution Analysis:** Visualizing the distribution of the risk score and other influent features, for clients whose loan applications are accepted versus those whose applications are denied.
- **Feature Importance and Insights:** Evaluating the features importance and their influence on the target, to better understand the determinants of loan approval and risk assessment.

By combining and comparing those techniques, methodologies and analysis, this project aims to provide a comprehensive understanding of the factors that influence loan decisions and how risk scores vary across different client groups. Initially, the models under comparison will be presented, and the results of their classifications and predictions will be reported. Following this, the focus will shift to the analysis of correlations between variables and other exploratory aspects of the dataset, through plots that favor the visualization of the data.

Pre-processing, model selection and validation

The dataset was found as an open-source resource and, fortunately, did not require extensive preprocessing, other than the removal of some duplicate elements. However, it contains both numerical features and features in string format, which may not be compatible with certain models. Therefore, it was necessary to perform an encoding operation on some columns using the *pandas* library with the *getdummies* method. To achieve better results, the data was shuffled to create a more varied and representative sample. This step was taken to avoid scenarios where entire portions of the dataset might consist solely of clients whose loan applications were rejected. Overall, this ensures a more balanced and reliable sample for analysis. Once completing the preprocessing phase, a large portion of the dataset (1500 rows out of 2000) was selected and appropriately split, first into features and target variables, and then into training and testing sets. The data were scaled using a *Standard Scaler* to normalize the feature values. This step ensures that all features are on a comparable scale. The resulting dataset was then prepared for the selection process using cross-validation. The models compared through cross-validation were a simple Decision Tree, a Random Forest, and a Support Vector Machine, all used as classifiers through *Scikit-Learn*, well known library in the realm of ML. During the selection phase, it was decided not to heavily adjust the hyper parameters but to maintain a basic configuration, postponing a more fine-tuned optimization (through *Grid Search*) only once the model that performed best during cross-validation had been identified. Assuming five folds, the results in terms of accuracy of the aforementioned techniques are shown in the tab below:

CLF	*Accuracy for each fold*					*Mean Accuracy*
Decision Tree	0,854	0,891	0,854	0,875	0,862	**0,8675**
Random Forest	0,93	0,9	0,9	0,916	0,912	**0,914**
SVM	0,95	0,916	0,937	0,937	0,92	**0,932**

Tab 1. Cross Validation scores in terms of accuracy

The Support Vector Machine classifier, which works by finding the hyperplane that best separates the data points into the predefined classes, achieved the best results during cross-validation. As a result, it was selected as the model to use. Subsequently, through *Grid Search CV*, a fine-tuning of its parameters was carried out to further optimize its performance. The grid explore different configurations changing *C*, *gamma* and the *kernel* of the SVM. *C* parameter controls the trade-off between achieving a low error on the training data and minimizing the margin size, which helps in generalizing better to unseen data. *Gamma* is a parameter that defines how much influence a single training data point has on the decision boundary (the hyperplane) that the SVM tries to create. The kernel defines the function used to transform the data into a higher-dimensional space, allowing the model to find a separating hyperplane in cases where the data is not linearly separable. The best set of hyper-parameters and the mean score in terms of accuracy related to them are:

- '*C*': **1**, '*gamma*': **1**, '*kernel*': '**linear**'
- *Best Mean Accuracy* : **0.967**

Training and Testing phase – Loan Approved Classification

Once data have been properly split and normalized, the training and testing phase begins, with 20% of the data reserved for testing. The predictions are then reported using a confusion matrix, and the performance scores, including accuracy, precision, and recall, are also provided to evaluate the model's effectiveness. The results obtained following the training and test phase on the first 1500 rows are reported in the tab below:

	Prediction	
Actual target	*Rejected*	*Approved*
Rejected	**225** (TN)	**6** (FP)
Approved	**6** (FN)	**63** (TP)

Accuracy:	**0,96**
Precision:	**0,91**
Recall:	**0,96**

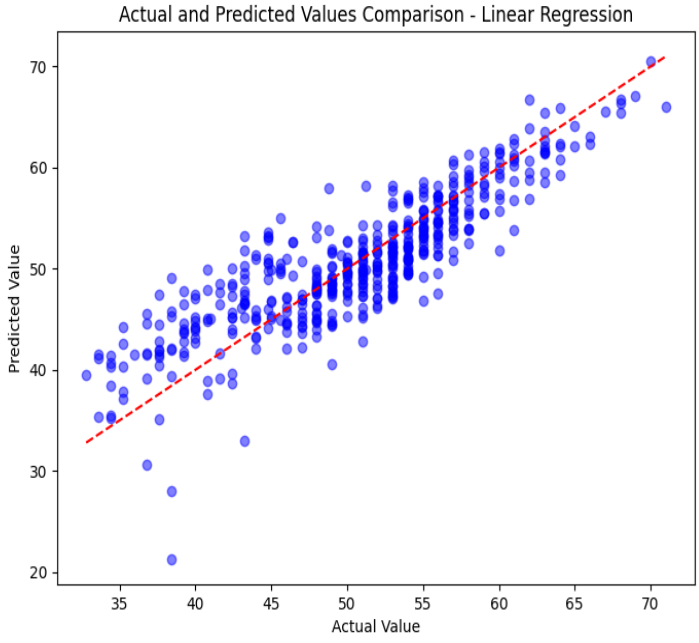
The model was then tested without any further training on the remaining portion of the dataset, consisting of 500 rows, in order to assess its ability to generalize. We can easily observe that the reported metrics did not experience a significant decline; in fact, they remain quite high. This suggests that the model has learned effectively to classify and generalize, demonstrating strong performance even on unseen data. The results are shown in the tab below:

	Prediction	
Actual target	*Rejected*	*Approved*
Rejected	**363** (TN)	**14** (FP)
Approved	**12** (FN)	**111** (TP)

Accuracy:	**0,948**
Precision:	**0,888**
Recall:	**0,948**

Training and Testing phase – Risk Score Regression

To avoid to overload the project, it was chosen to predict the risk score associated with each client using linear regression, without comparing different models. Dealing with continuous values, scaling is crucial, so once have normalized data, the regression model was trained and tested and its performance were measured in terms of root mean squared error (RMSE) and R^2 Score. Again the model was eventually tested on unseen data to analyze the ability of generalize.



Train and Test	
RMSE	**3,84**
R^2	**0,756**
Unseen Data	
RMSE	**3,74**
R^2	**0,742**

ID	Real Value	Pred Value
1500	53	51,56
1501	56	54,93
1502	55	51,78
1503	55	57,66
1504	53	51,73
1505	68	66,71
1506	53	50,46

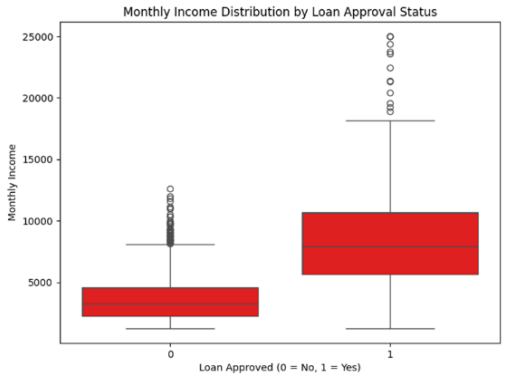
The results of the linear regression in predicting the risk score are quite acceptable. A Lasso regression was also implemented, but it did not led to significant improvements. In the graph, we can observe that the points representing the actual and predicted values are relatively close to the regression line, indicating a reasonable level of accuracy. Presumably, better optimization and tuning, such introducing polynomial features to capture non-linear relation, or the usage of different regressors (KNN, Perceptron, MLP), would have likely led to improved results.

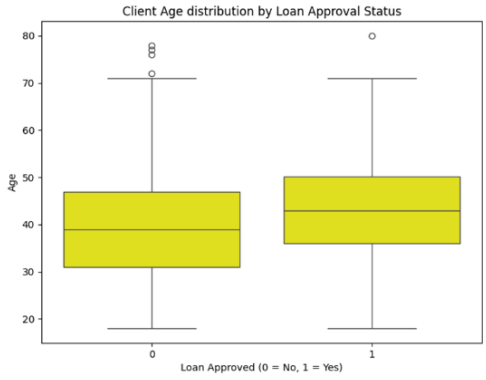
Correlation, data distribution and insights

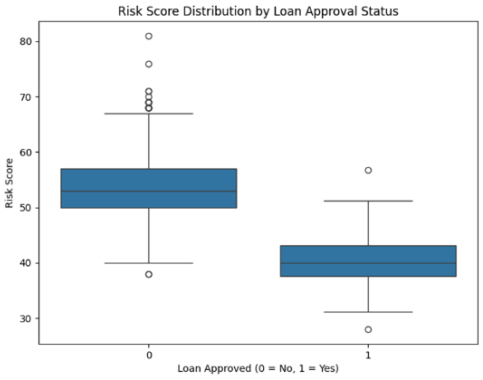
To gain insights into the correlation between features and the target, and to analyze which features most influence the approval or denial of a loan, the decision was made to calculate Pearson correlation coefficients. These coefficients range from -1 to 1 and measure the strength and direction of the linear relationship between each feature and the target. The obtained results are aggregated in the tab below:

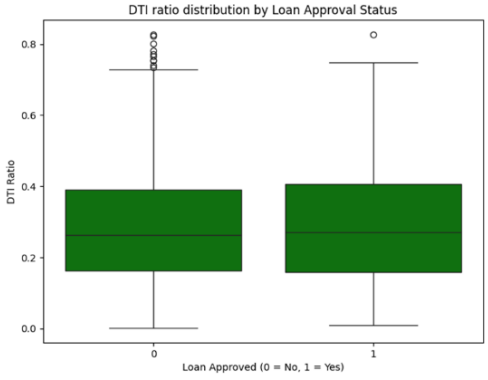
Positive Linear Correlation		***Negative Linear Correlation***	
MonthlyIncome	**0,613**	SavingsAccountBalance	-0,096
AnnualIncome	**0,612**	LoanDuration	-0,106
Age	0,146	MonthlyDebtPayments	-0,113
Experience	0,136	EducationLevel_HighSchool	-0,165
EducationLevel_Doctorate	0,129	MonthlyLoanPayment	-0,201
CreditScore	0,120	BaseInterestRate	-0,231
NetWorth	0,109	LoanAmount	-0,258
TotalAssets	0,103	TotalDebtToIncomeRatio	**-0,420**
LengthOfCreditHistory	0,102	RiskScore	**-0,749**

The features/attributes that most positively influence the likelihood of receiving a loan appear to be the monthly/annual income and the age of the client (although with weak correlation), while those that most negatively influence it are the total debt to income ratio and of course the risk score, even the loan amount and the interest rate play a bit of a role. It is possible to show, by means of boxplot, the distribution of these data in relation to the fact that the loan was approved or not.









As expected, the distributions of risk score and monthly income, features with higher Pearson correlation coefficients (strongly correlating with loan approval, one positively and the other negatively), are more distinctly separated between the two scenarios. These features exhibit a clearer separation compared to others with weaker correlations, such as age or DTI ratio. Although for risk score and monthly income we can observe a clear separation between the median values of the two scenarios—suggesting a threshold above which a loan is more likely to be granted—there are still many outliers. This indicates that for those clients, other features likely had a greater influence. We can attempt to explore the reasons behind some of these outliers. For instance, we can observe a customer whose loan was approved despite having a Risk Score higher than the median value of those whose loans were not approved. By analyzing some of the customer's features, we can notice highly favorable values for Annual Income and DTI Ratio, along with a very modest Loan Amount in respect to his Net Worth:

ID	Annual Income	DTI Ratio	Age	Loan Amount	NetWorth
97	283369	0,23	43	9851	46346

In fact, these features are among those most strongly correlated with the target. This explains why, despite a less-than-ideal risk score, the customer's loan was approved. Following the same reasoning we can notice several outliers in the monthly income data distribution. A large number of costumers did not received a loan despite having a monthly income greater than the median value of the ones who received it. In this case, given the presence of multiple heterogeneous data points, it becomes more challenging to pinpoint the features that negatively influenced the likelihood of loan approval, as there are numerous factors involved. A subset of these features is shown in the table below:

ID	Interest Rate	DTI Ratio	Age	LoanAmount	NetWorth
6	0,30	0,68	58	32748	9086
13	0,26	0,11	18	28309	21123
60	0,26	0,36	34	10095	3647
227	0,23	0,31	27	16753	2430
228	0,26	0,29	31	31780	3907
306	0,20	0,45	55	31183	43869
352	0,25	0,48	50	35730	70708
474	0,24	0,18	59	29612	6794
491	0,28	0,20	57	22818	113614
498	0,26	0,33	29	24816	7333
503	0,22	0,13	46	31216	7524
548	0,25	0,71	44	42536	13133

For some customers, a young age might have played a role, while for others, the debt-to-income ratio could have been a determining factor. For yet others, the loan amount might have been deemed too high relative to what the bank considers affordable for the customer. It is not possible to provide a single, definitive explanation. In conclusion, we can say that, generally, the applied models showed acceptable performance, particularly for classification tasks, and that the features identified as most influential exhibit distinct distributions, assessing their discriminant behavior. Certainly, employing different techniques or conducting a more in-depth analysis of the outliers and, more generally, of the attributes in the dataset, could lead to even more precise and detailed results, providing an even clearer understanding of the factors that influence the likelihood of receiving a loan.