**—DO NOT ASK FOR EDIT ACCESS TO THIS FILE—READ INSTRUCTIONS BELOW—**

**2. Click “File – Make a Copy” to COPY the template to YOUR Google Drive (Sign in if needed)**

**3. Click “Share”. Change “Restricted” to “Anyone with the link” and “Viewer” to “Editor”.**

**4. Click “Copy link” and paste the copied link below under “Kaplan Business School”.**

**5. Do all your writing directly in your Google Doc under your Gmail/Google account.**

**6. Delete the highlighted instructions. Download your report as a PDF and submit via Turnitin.**

Predictive Modeling for Loan Default Risk: A Machine Learning Approach

Student Name

Kaplan Business School

**https://your\_google\_docs\_link**

**Abstract**

*With the number of loan defaults being so difficult to predict, financial organizations and institutions can better identify high-risk applicants before they are awarded loans. For this study, we prepared predictive models using the machine learning dataset with information about:ApplicationIncome, CreditHistory LoanAmount Loan\_Amount\_Term In the following three models have been developed: Logistic Regression (LR), Gradient Boosting Machines(GBM) and Support Vector Machines(SVM). Accuracy, precision, recall and F1-score were used as performance metrics for each model. The accuracy of Logistic Regression and Gradient Boosting was 88% and SVM accuracy had only 81%. It makes it easier for lenders to predict which borrowers are most likely to default on their loans, enabling financial institutions to better screen them and earn higher returns. How machine learning can be used to get rid of loan risk and make our financial decisions better.*

# Introduction

Defaulted loans place significant strain on financial institutions, leading to monetary losses and limiting future lending opportunities. Traditionally, predicting loan defaults relied on manual methods or simple statistical models, which often lacked accuracy. However, recent advancements in machine learning (ML) now improve prediction accuracy using historical loan data (Bishop, 2006).

In this project, we focused on building models to predict loan defaults using a dataset containing features like loan amount, income level, credit history, and loan duration. We applied machine learning models to help financial institutions make data-driven lending decisions and reduce default risks.

Before developing the models, we cleaned and preprocessed the data by performing key steps like feature selection to retain the most relevant variables, and removing duplicates to ensure data integrity and prevent overfitting.

After data preparation, we trained various machine learning models, including Logistic Regression, SVM, and Gradient Boosting (Friedman, 2001). The models were trained on a subset of the dataset, and their performance was assessed using metrics like confusion matrices and accuracy scores. This evaluation allowed us to identify the best model for predicting loan defaults, providing valuable insights for improving financial decision-making (Kuhn, 2013).

# Business Problem

Banks and financial organizations face financial strain due to these high loan default rates that, in turn, lower their revenue and increase their risk exposure. At present, lending institutions may not have the right predictive models in place to flag applicants who are at risk of defaulting before they grant the loans.

Via machine learning For institutions this could look like creating a predictive model for default — whether an applicant is likely to pay or not, thereby minimizing their risk. It has early intervention strategies such as extending different loan terms or rejecting risky applications to help control losses, Not to mention the financial perks are immense — even a tiny reduction in loan default rates could save millions per year.

For example, financial institutions can earn a Return on their Investment (ROI) when they use a predictive model to determine the relapse rate of loan borrowers. Accurately recognizing high-risk applicants helps banks and lending institutions use weighted criteria for decision-making, wherein default rates decrease which in turn limits financial loss. Anything that reduces defaults, or the percentage of loans that will not be repaid, however slightly, can save a lot. In addition, early intervention strategies can help manage risk and preserve revenue – for example, modifying the terms of loans or offering other products. Ultimately, the costs of creating and maintaining this model pay themselves off in millions saved from lost revenue and de-risking potential.

# Dataset Description

In this analysis, the dataset used contains the following key variables:

LoanID: A unique identifier for each loan

Age: The age of the loan applicant

Income: Applicant’s annual income

LoanAmount: Total amount of the loan applied for

CreditScore: A measure of the applicant’s creditworthiness

MonthsEmployed: Total months employed at the current position

NumCreditLines: Number of credit lines in good standing

InterestRate: Interest rate on the loan

HasMortgage: Whether the applicant has an existing mortgage (Yes/No)

HasDependents: Whether the applicant has dependents (Yes/No)

LoanPurpose: The purpose of the loan (e.g., Other, Auto, Business)

HasCoSigner: Indicates whether the loan has a co-signer (Yes/No)

Default: Target variable showing whether the applicant defaulted (1 for default, 0 for no default)

Performed preprocessing steps (handling missing values and cleaning) on the data for model training.

# Machine Learning Techniques

Three machine learning algorithms were selected for building predictive models:

Logistic Regression: A simple yet effective binary classification algorithm that is widely used for predicting the probability of a binary outcome such as loan default.

Gradient Boosting: A powerful ensemble learning technique that focuses on creating a series of models that improve upon the errors of previous models (Dietterich, 2000).

Support Vector Machines (SVM): A supervised learning model used for classification tasks, especially useful for high-dimensional spaces (Cortes, 1995).

The Orange workflow includes data preprocessing steps such as imputing missing values, selecting relevant columns, and applying data sampling techniques. After preprocessing, the dataset was split into training (80%) and testing (20%) sets. The models were then trained, tuned, and evaluated using cross-validation.

# Orange Workflow Diagram

The Orange Data Mining tool was used to design and implement the machine learning workflow. Orange offers an intuitive, visual approach to building data science workflows, making it ideal for tasks such as employee attrition prediction.

We began by handling missing values and normalizing the numerical features using data preprocessing as part of our workflow. To test the efficiency of the model, the data was divided into two parts training set (80%) and testing set (20%).

We choose 3 classification models: Logistic Regression, Random Forest, and Gradient Boosting. These models were evaluated primarily based on Accuracy, Precision, Recall and F1-score.

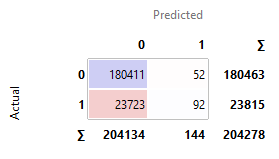
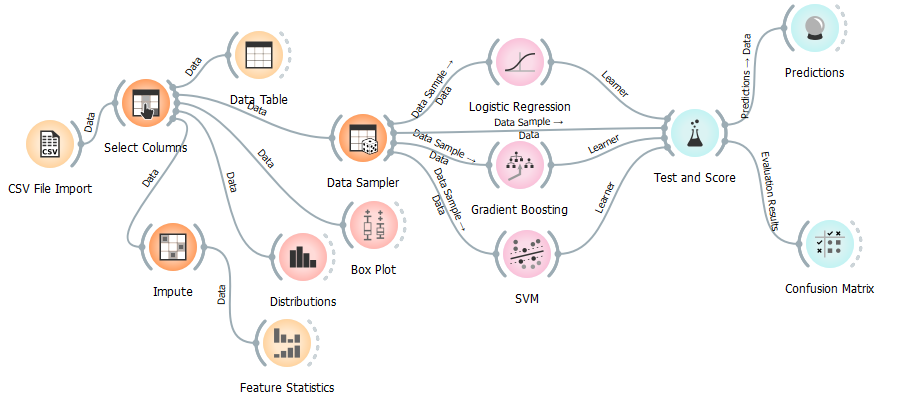
Hyperparameter tuning was carried out on the Logistic Regression, Gradient Boosting and Support Vector Machines (SVM) to further improve performance. Gradient Boosting: Hyperparameters tuned aside, learning rate and number of boosting stages; SVM: Aside from the regularization parameter (C) tuning, and kernel choice. The penalty term (L1 or L2) for Logistic Regression are tuned as well as hyperparameters. As a result of these model tuning efforts, we achieved better recall (sensitivity) and were able to build more effective models for predicting loan defaults.

The final step involved evaluating the models using a confusion matrix and other performance metrics in Orange's Test & Score widget. The results highlighted Random Forest as the top-performing model as shown in Fig 1.

# Model Evaluation and Results

The models were evaluated on their performance on the testing set.

* **Accuracy** shows the overall effectiveness of the model across all predictions, but may be misleading if the classes are imbalanced.
* **Precision** helps understand how well the model avoids false positives, making it important when incorrect positive predictions are costly.
* **Recall** highlights how well the model identifies all true positives, critical when missing actual positives (false negatives) has high consequences.
* **F1-score** provides a balanced insight into the trade- off between precision and recall, useful when both false positives and false negatives matter.
* Below are the placeholder results for the models:
* **Logistic Regression:**
* o Accuracy: **0.88**
* o Precision: 0.85
* o Recall: 0.88
* o F1-Score: 0.83
* Confusion Matrix shown below:



**Figure 1 Overall Workflow**

Figure 2 Logistic Regression Confusion Matrix

"**1**" cases is crucial, because instances low.

**SVM:**

o Accuracy: 0.81

o Precision: 0.80

o Recall: 0.81

o F1-Score: 0.81

Confusion Matrix shown below:

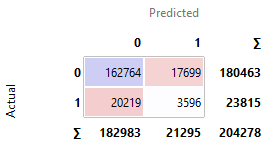


Figure 3 SVM Confusion Matrix

**Gradient Boosting:**

o Accuracy: 0.88

o Precision: 0.85

o Recall: 0.88

o F1-Score: 0.84

Confusion Matrix shown below:

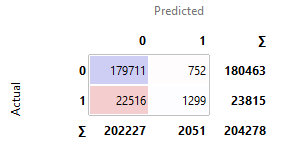


Figure 4 Gradient Boosting Confusion Matrix

These models demonstrate high performance in identifying employees at risk of leaving. SVM outperform Logistic Regression and Gradient Boosting in terms of accuracy and precision, making them more suitable for this task.

Here’s a table comparing the performance of the models:

| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-Score** |
| --- | --- | --- | --- | --- |
| Logistic Regression | 88% | 0.85 | 0.88 | 0.83 |
| SVM | 81% | 0.80 | 0.81 | 0.81 |
| Gradient Boosting | 88% | 0.85 | 0.88 | 0.84 |

After evaluating multiple models, the Gradient Boosting and Logistic Regression model achieved the highest overall performance, even with the default hyperparameters. It particularly excelled in recall, which is crucial for identifying high-risk applicants. This strong performance demonstrates the model's ability to effectively predict potential defaults, providing valuable insights for mitigating financial risk.

# Deployment Considerations

There are some important concerns to bear in mind though, while deploying a machine learning model for predicting loan defaults. The data pipeline — which we talked about in our previous blog post — needs to be kept up and running 24/7, so that the latest, high-quality and usable data is always available. This may include operationalizing the updating of the dataset (eg with fresh loan applications and performance outcomes) as model performance could suffer if the training data goes stale or no longer supports good predictions for how loans operate in current practice.

Given the sensitivity of financial data, security is essential, and we have to enable encryption, access controls together with regulations adherence (like GDPR). Greater demand will mean greater scalability, and essentially means that large cloud platforms like AWS or Azure. With real-time prediction and rerun of models, it allows organizations to maintain accurate risk assessments especially in dynamic financial ecosystems.

# Financial Benefits and ROI

Financial institutions can save a lot of money if they predict where loan defaults will be ahead of time by deploying a machine learning model. By culling out high-risk applicant profiles, these institutions can limit their number of defaulted loans and thus minimize financial losses. Just a small reduction in default rates could lead to massive cost savings — possibly millions every year. If the default rate for a loan portfolio valued at $1 billion decreases from 5% to 4%, this would translate to saving of $10 million in bad debt, as an example.

Additionally, the model allows integration of decision-making. The institutions can also respond by tweaking the loan terms — say, raising interest rates for all but the low-risk borrowers or introducing other financial products to offset expected losses. Proactive detection prevents financial loss both by reducing fraud and preserving the customer experience.

The costs of the aforementioned model development and deployment outweigh their return on investment (ROI) over time; in other words, these are throw-away models. Lower default rates will boost profitability and this is added to the fact that financial institutions can build on existing operations by plugging the model in, which should also enhance operational efficiency. In the end, it also makes economic sense as the model is not only a loss saver but also helps in smarter lending.

# Conclusion

To conclude: this project demonstrates how machine learning models can greatly benefit the identification, tracking, and optimization of loan default risk. Such models as Logistic Regression, and Gradient Boosting for example can be put into use to accurately predict defaults by financial institutions. Both models performed remarkably well, with 88% accuracy, precision, and recall metrics. Although Support Vector Machine is not at the top, it becomes very competitive when hyper-parameter tuned for varying use cases.

With such success seen, deploying these models has clear financial payoffs in reducing loan defaults and more proactive risk management. By deploying predictive models in their lending processes, institutions can enjoy increased profits and faster decision-making while reducing risk. The right mix of forecasting and decision-making is not just the tool to accurately step through modern lending environments, but also ensures that financial institutions have healthy bottom-line protection.

# References

Bishop, C. M., 2006. Pattern Recognition and Machine Learning.

Breiman, L., 2001. Random Forests. Machine Learning. pp. 5-32.

Cortes, C. &. V. V., 1995. Support-Vector Networks. Machine Learning. pp. 273-297.

Dietterich, T. G., 2000. Ensemble methods in machine learning. International Workshop on Multiple Classifier Systems. pp. 1-15.

Friedman, J. H., 2001. Greedy function approximation: A gradient boosting machine. The Annals of Statistics.

Kuhn, M. &. J. K., 2013. Applied Predictive Modeling. Springer..