Machine Learning for Economists

Final Project

“PCA vs Bayesian Shrinkage vs OLS and PLS:
Predicting Loan Approval through Dimensionality Reduction and Regularization”

Santiago Viola

Moira Patricia Clavin

Maria Catalina Avaca

Pedro Straface

Paper Summary

"Forecasting Using a Large Number of Predictors: Is Bayesian Shrinkage a Valid Alternative to Principal Components?"
De Mol, Giannone, and Reichlin (2008)

This paper talks about the problem of forecasting when the number of potential predictors is too large relative to the number of observations, this situation is commonly referred to as the curse of dimensionality. The authors propose to use the Bayesian regression with shrinkage priors as an alternative to the widely used Principal Component Regression (PCR) approach in macroeconomic forecasting.

Specifically, they consider two types of priors on the regression coefficients:

1. A Gaussian (normal) prior, which leads to a Ridge regression estimator
2. A double-exponential (Laplace) prior, which leads to a Lasso regression estimator.

Under the Gaussian prior, the posterior mode is equivalent to minimizing a penalized least squares function with an L2 norm penalty:

$$\min_{\beta}(y - X\beta)'(y - X\beta) + \lambda \sum_{j} \beta_j^2,$$

On the other hand, the double-exponential prior leads to an L1 penalized least squares problem, which performs both parameter shrinkage and variable selection.

$$\min_{\beta}(y - X\beta)'(y - X\beta) + \lambda \sum_{j} |\beta_j|,$$

The paper uses the large U.S. macroeconomic dataset from Stock and Watson (2002) to compare the forecasting performance of Bayesian regression (under both priors) with that of principal component forecasts. The results show that the forecasts obtained from Bayesian Ridge and Lasso regressions are highly correlated with those produced by principal component methods and achieve similar mean-square forecast errors (MSFE) across a broad range of prior choices. Therefore, Bayesian shrinkage is shown to be a valid and robust alternative to PCA based forecasting.

They prove that forecast consistency under a Gaussian prior is achieved provided that the degree of shrinkage increases with the number of predictors. When the data follow an approximate factor structure, Bayesian regression and principal component methods asymptotically converge, since both give higher weight to the directions in the data associated with the dominant eigenvalues of the covariance matrix.Finally, the paper notes that while Lasso regression produces more scattered models through variable selection, its forecasts are not better than those obtained from Ridge or PCA. This can be explained by the strong multicollinearity present in macroeconomic panels, where a few variables can summarize most of the common variation. Therefore, Bayesian shrinkage approaches, especially the Gaussian prior, provide a consistent and efficient framework for forecasting with large panels of time series.

1 - Motivation

The motivation for this study originates from our interest in exploring dimensionality reduction and regularization techniques in the field of forecasting and predictive modeling. The paper “Forecasting Using a Large Number of Predictors: Is Bayesian Shrinkage a Valid Alternative to Principal Components?” by De Mol, Giannone, and Reichlin (2008) presents an appealing framework that directly compares Principal Component Analysis (PCA), a method we were already familiar with, to Bayesian shrinkage approaches such as Ridge and Lasso regressions, which were less known to us prior to this work.

Our study extends the comparison by also incorporating Ordinary Least Squares (OLS) as a benchmark model and Partial Least Squares (PLS) as an additional supervised dimensionality reduction technique. This perspective allows us to evaluate not only how shrinkage and factor-based methods perform relative to each other but also how they compare to the traditional linear regression approach and a method that combines the predictive orientation of shrinkage with the dimensionality reduction of PCA.

This comparison provides an opportunity to improve our understanding of how different methods handle the curse of dimensionality, the situation in which the number of predictors is large relative to the number of observations, leading to instability and overfitting in conventional econometric models. While PCA reduces dimensionality by extracting components that summarize common variation across variables, PLS looks for components that maximize covariance with the target variable and shrinkage methods address the same issue by imposing penalties that constrain the size of regression coefficients, improving generalization ability.

We found this topic particularly relevant for Machine Learning for Economists, as it connects traditional econometric modeling (OLS and PLS) with modern machine learning techniques (Ridge and Lasso) offering a comparative view between unsupervised dimensionality reduction (PCA) and supervised regularization. Understanding how these methods differ in theory and forecasting performance contributes to building intuition about how machine learning can improve empirical economic analysis.

To complement the theoretical discussion held in the paper previously named, we extend their framework to a different context: loan approval prediction. The Loan Approval dataset contains a variety of financial and demographic variables that in conjunction determine whether a loan application is approved or rejected. This type of dataset reflects a realistic high dimensional setting where multiple correlated predictors influence a binary decision outcome.

In this sense, the loan approval problem provides a natural application for comparing PCA and shrinkage based methods. It allows us to evaluate how each technique performs in terms of predictive accuracy, interpretability, and robustness when faced with multicollinearity and heterogeneous sources of information. Also, this exercise demonstrates how the methodological insights from De Mol, Giannone, and Reichlin, initially developed for macroeconomic forecasting, can be effectively applied to classification problems in financial economics.

2 - Introduction

This work focuses on the comparison of different linear prediction methods designed to address the problem of high dimensionality in econometric and machine learning applications. Following the framework proposed by De Mol, Giannone, and Reichlin, we aim to evaluate the performance of Bayesian shrinkage methods in relation to PCA and OLS. Additionally, we consider PLS a supervised dimensionality-reduction method that, unlike PCA, extracts components by maximizing the covariance between predictors and the dependent variable.

The main objective is to examine how these approaches handle situations in which the number of predictors is large relative to the number of observations, generating collinearity and potential overfitting.OLS serves as the baseline model in this comparison, providing a reference for assessing how regularization and dimensionality reduction techniques improve upon the limitations of standard linear regression under high-dimensional settings.

PCA represents a dimensionality reduction technique, summarizing correlated predictors into a smaller set of orthogonal components that capture most of the variance in the data. PLS, in contrast, seeks latent components that are directly relevant for prediction, potentially improving forecasting performance when the relationship between XXX and YYY is strong but obscured by multicollinearity. In this framework, PLS can be viewed as an intermediate approach that combines the dimensionality reduction of PCA with the predictive orientation of shrinkage techniques, thus bridging unsupervised and regularized methods.

Bayesian shrinkage applies regularization through previous distributions on the regression coefficients, typically Gaussian or double exponential, producing Ridge-type and Lasso-type estimators respectively. These techniques penalize large coefficients improving generalization while maintaining predictive accuracy.

We apply these methods to the loan approval prediction problem, a relevant real-world context in which multiple correlated financial and demographic variables jointly determine whether a loan is approved. This dataset provides an appropriate testing ground for evaluating how each method manages multicollinearity, bias–variance trade-offs, and predictive performance.

By comparing PCA, PLS, Ridge and Lasso (Bayesian shrinkage) and OLS, we aim to determine which approach achieves the best balance between accuracy, interpretability, and robustness. Ultimately, this study looks to demonstrate that Bayesian shrinkage provides a valid and flexible alternative to traditional dimensionality reduction and least-squares methods in predictive economic modeling.

3 - Methods

The methodological approach of this work builds on the ideas presented by *De Mol, Giannone, and Reichlin (2008)* and on the theoretical foundations discussed in the course *Machine Learning for Economists*. The objective is to evaluate and compare different techniques used for prediction in high-dimensional settings , situations where the number of predictors is large relative to the number of observations, which makes traditional estimation methods unstable and prone to overfitting.

The starting point of our analysis is the classical linear model, in which a dependent variable (in this case, the loan approval decision) is expressed as a linear combination of several independent predictors (financial and demographic characteristics of each applicant). However, when these predictors are highly correlated or numerous, the OLS estimator becomes inefficient, small variations in the data may lead to large fluctuations in the estimated coefficients, reducing the model's ability to generalize to new observations.

Principal Component Analysis (PCA) offers one way to address this problem. Rather than using the original correlated variables, PCA transforms them into a smaller number of uncorrelated components that summarize most of the variation in the dataset. In this study, the number of components is selected according to the cumulative explained variance criterion, retaining enough components to capture the majority of the variance while discarding those that contribute little information. In this case, the criterion indicated the use of three components, which together account for most of the variability in the predictors. By keeping only these components, the model becomes more stable and avoids the noise generated by redundant predictors. Still, this approach may come at the cost of interpretability, since principal components are linear combinations of all original variables and do not have a straightforward economic meaning.

PLS provides a complementary strategy. While PCA focuses exclusively on explaining the variance of the predictors (X), PLS constructs latent components that maximize the covariance between pax and the dependent variable Y. This makes PLS a supervised dimensionality-reduction method, directly oriented towards improving predictive accuracy. In practice, PLS can outperform PCA when the relationship between predictors and the target variable is strong but obscured by multicollinearity, effectively combining the dimensionality reduction of PCA with the predictive focus of regularized models.

The Bayesian shrinkage approach follows a different strategy. Instead of reducing the number of variables, it keeps all predictors but imposes a constraint that "shrinks" their estimated coefficients toward zero. This shrinkage is implemented through prior distributions that penalize large coefficients. A Gaussian prior leads to a smoother penalization (as in Ridge regression), which keeps all predictors active but reduces their influence, while a double-exponential prior (as in Lasso regression) induces sparsity by pushing some coefficients exactly to zero, effectively performing variable selection. Both methods aim to increase stability and prevent overfitting, with the difference that Ridge focuses on regularization and Lasso also promotes interpretability by identifying the most relevant variables.

Alongside these methods, we include OLS as a benchmark model. OLS serves as the baseline for comparison: it provides unbiased estimates but is highly sensitive to collinearity and noise, which can lead to unstable predictions in high-dimensional settings. Together, these techniques form a continuum from unregularized estimation (OLS) to full regularization (Lasso), allowing us to analyze how predictive performance evolves as regularization strength increases.

The empirical application uses the Loan Approval dataset, which contains information on loan applicants' income, assets, credit history (CIBIL score), loan amount, employment status, and other demographic variables. The dependent variable indicates whether the loan was approved, making this a binary prediction problem. The dataset presents typical challenges of real-world financial data: correlated variables, heterogeneous scales, and potential redundancy among predictors. These characteristics make it an appropriate setting to test the different methodologies under conditions of high dimensionality and multicollinearity.

All models are implemented and tested in a Jupyter Notebook (.ipynb) file. This computational environment allows us to clearly present the practical steps of the analysis, including data preprocessing, model estimation, and performance evaluation. In particular, we will:

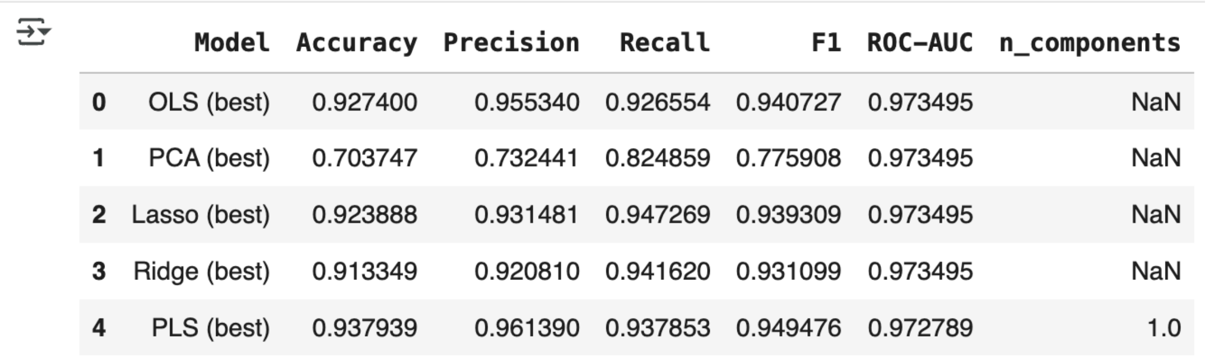
- Apply each method (OLS, PCA, PLS, Ridge and Lasso) to the same training and testing samples of the dataset.
- Use cross-validation techniques to select optimal parameters (such as the number of components for PCA and PLS, and the regularization parameter λ for Ridge and Lasso).
- Predictive performance will be evaluated and compared using standard classification metrics, including accuracy, precision, recall, and the F1-score, together with the confusion matrix, which provides a detailed view of correctly and incorrectly classified observations.
- Visualize the results through coefficient paths, variance explained plots, and error comparison tables.

This procedure allows us to assess not only which model achieves the best predictive performance, but also how each technique affects interpretability and robustness. Following the reasoning of De Mol et al. (2008), we expect that PCA, PLS and Bayesian shrinkage will perform similarly in predictive terms, with shrinkage offering a more flexible structure that can adapt to the degree of correlation among predictors. By applying these models to a real-world dataset outside the macroeconomic context of the original paper, we aim to demonstrate the general applicability of these techniques to other fields, in this case, financial decision-making, where prediction accuracy and model transparency are both essential.

4 - Results

The comparison between OLS, Ridge, Lasso, Partial Least Squares (PLS), and PCA provides clear evidence about the trade-offs between model complexity, interpretability, and predictive accuracy in high-dimensional settings.
Across all performance metrics (Accuracy, Precision, Recall, F1-score, and ROC-AUC) the results consistently indicate that regularized methods outperform unregularized and dimensionality-reduction approaches.

All models were estimated in a binary classification setting, where the target variable represents the loan approval decision. Logistic regression was used for OLS, Lasso, and PCA-based models. Ridge and PLS, however, were implemented in their linear forms (as logistic versions are not directly available) producing continuous fitted values interpreted as probabilities, which were then converted into binary classifications using a 0.5 threshold to ensure comparability across methods while preserving each model's estimation framework



	Model	Accuracy	Precision	Recall	F1	ROC-AUC	n_components
0	OLS (best)	0.927400	0.955340	0.926554	0.940727	0.973495	NaN
1	PCA (best)	0.703747	0.732441	0.824859	0.775908	0.973495	NaN
2	Lasso (best)	0.923888	0.931481	0.947269	0.939309	0.973495	NaN
3	Ridge (best)	0.913349	0.920810	0.941620	0.931099	0.973495	NaN
4	PLS (best)	0.937939	0.961390	0.937853	0.949476	0.972789	1.0

Figure 1. Model performance comparison across methods.

Among all models, as shown in Figure 1, PLS regression achieved the highest overall performance, with an accuracy of 0.938 and an F1-score of 0.949, slightly exceeding Ridge and OLS, which both reached an AUC of 0.973. This strong performance reflects PLS's hybrid nature: by constructing latent components that maximize the covariance between predictors and the target variable, it retains the interpret ability of linear models while filtering out noise and irrelevant variance.

In this sense, PLS can be seen as an intermediate approach between shrinkage (Ridge/Lasso) and unsupervised reduction (PCA). The optimal PLS configuration required only one latent component, indicating that most of the relevant predicting information can be captured along a single dimension combining the most informative predictors. This efficiency further supports the idea that PLS achieves effective dimensionality reduction without compromising predictive accuracy. Ridge and Lasso displayed very similar results, suggesting that the addition of shrinkage penalties substantially improves stability and predictive power relative to OLS, while maintaining strong generalization ability.

The performance metrics remain consistent across validation folds, this confirms that that regularization enhances model robustness and prevents overfitting. Lasso's slightly lower recall but comparable precision shows its ability to eliminate less informative predictors without compromising classification accuracy.

By contrast, PCA regression performed notably worse, with an accuracy of 0.70 and an F1-score of 0.78, reflecting its limitation in capturing the predictive structure of the data when the principal components explaining variance are not necessarily the most relevant for prediction. This confirms one of the key insights from *De Mol, Giannone & Reichlin (2008)* — that while PCA efficiently summarizes information, Bayesian or penalized shrinkage can better adapt to the specific predictive relationships between variables and outcomes.

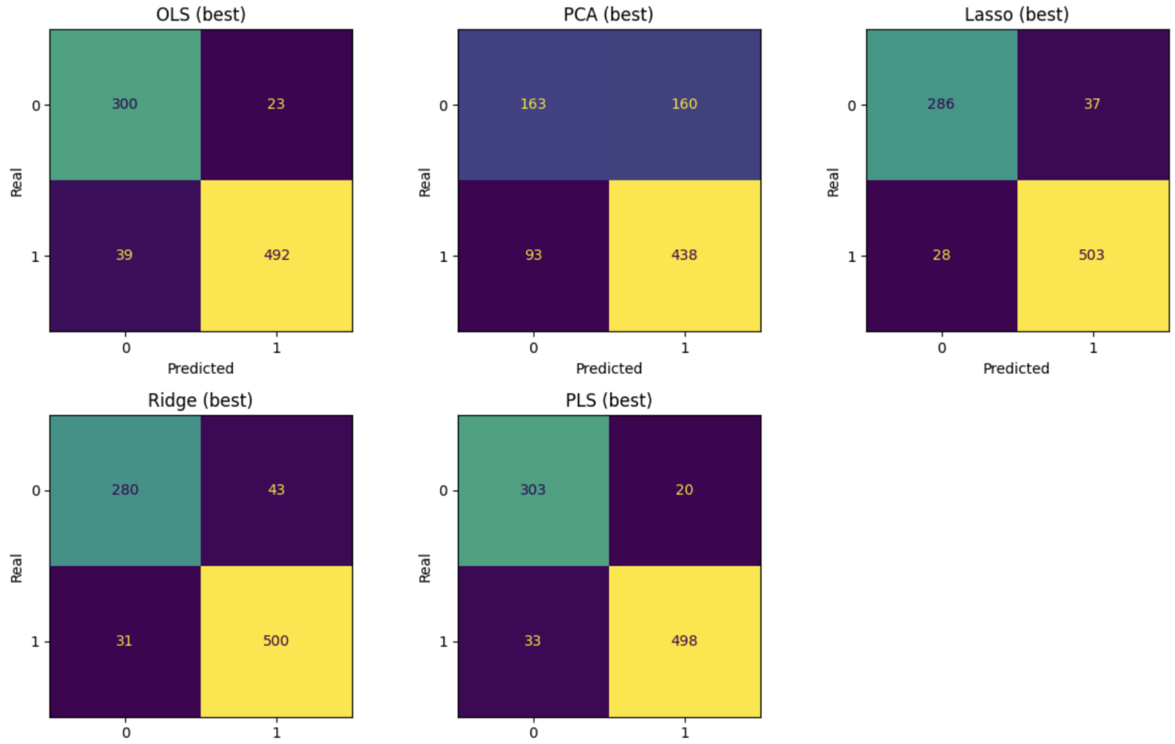


Figure 2. Confusion matrices for the five estimated models.

As shown in Figure 2, the confusion matrices confirm these findings visually: Ridge, Lasso, OLS, and PLS correctly classify over 93% of positive cases, while PCA shows a significantly higher rate of false positives and false negatives, revealing its lower discriminative capacity. Similarly, the ROC curves of Ridge, Lasso, and PLS almost overlap, confirming their near-identical discrimination power, whereas PCA's curve lies distinctly below, evidencing its inferior sensitivity.

Overall, the inclusion of regularization terms (particularly in Ridge and Lasso) yields more robust and reliable models, capable of managing multicollinearity and overfitting without relying on arbitrary dimension reduction. The superior performance of PLS reinforces our conclusion, showing that predictive dimensionality reduction guided by the response variable can rival or even surpass classical shrinkage estimators in structured, high-dimensional problems. These empirical results, derived from the Colab implementation, also provide clear evidence supporting De Mol, Giannone & Reichlin's (2008) theoretical claims: shrinkage and supervised reduction strategies deliver forecasts that remain stable, efficient, and interpretable when the number of predictors is large relative to the sample size.

5 - Discussion

The results obtained in this analysis provide clear evidence on the advantages of shrinkage-based methods in high-dimensional prediction problems. However, these results should be interpreted beyond their numerical performance. Each technique presents trade-offs between accuracy, interpretability, and computational simplicity that are relevant depending on the empirical context.

All models were estimated for a binary classification task, where the target variable represents the loan approval decision. Logistic specifications were used for OLS, Lasso, and PCA-based models, while Ridge and PLS were implemented in their linear forms due to the unavailability of a logistic alternative. Their continuous fitted values were transformed into probabilities and classified using a 0.5 threshold. This design ensures comparability across models while preserving their individual estimation logic.

In the case of the loan approval dataset, Ridge and Lasso regression showed superior predictive accuracy, confirming their ability to manage correlated variables and prevent overfitting. Yet, their main strength lies in stability rather than interpretability. Ridge keeps all predictors in the model, which makes it difficult to identify the most influential variables, while Lasso improves this aspect by shrinking some coefficients to zero, but may discard variables that are weakly correlated yet still economically relevant.

PLS deserves a special mention because it combines elements of both shrinkage and dimensionality reduction. By extracting latent components that maximize the covariance between predictors and the response variable, PLS captures the predictive structure of the data more effectively than PCA. In this application, the optimal PLS model required only one latent component, achieving the highest accuracy and F1-score among all approaches. Its advantage lies in translating complex predictor interactions into a few interpretable latent factors directly linked to the outcome, which provides a balance between parsimony and explanatory power.

Principal Component Analysis, on the other hand, performed worse in predictive terms but remains a useful dimensionality-reduction tool when the goal is to summarize the structure of information rather than make precise forecasts. Its main limitation is that principal components are abstract and do not have a straightforward economic meaning, which complicates their interpretation in applied policy or financial settings.

From a methodological perspective, the comparison across these models illustrates a broader point made by De Mol, Giannone, and Reichlin (2008): there is no universally superior technique, but rather methods that are better suited to specific data environments. In macroeconomic forecasting, where underlying factors drive co-movement among many variables, PCA can perform similarly to shrinkage. In financial microdata, where variable relationships are often more irregular, Bayesian shrinkage usually provides a better fit. PLS stands somewhere in the middle: it behaves like a shrinkage method in how it focuses on prediction, yet it keeps the factor-model idea of reducing dimensions. This mix probably helps it perform well in datasets that are structured but not entirely factor-based.

Another aspect worth noting is the computational simplicity of modern shrinkage estimators. Regularization techniques such as Ridge and Lasso can be implemented efficiently even in large datasets, making them particularly appealing for real-world applications in economics and finance, where interpretability and speed are equally important. PLS also remains computationally efficient, since it reduces the predictor space early in estimation, offering a scalable alternative for predictive modeling when interpretability and performance must coexist.

Finally, this exercise also shows how the integration of traditional econometric thinking with modern machine learning tools can improve empirical analysis. Instead of viewing these approaches as competing paradigms, they can be seen as complementary. Dimensionality reduction, shrinkage, and cross-validation together provide a framework that enhances both the predictive accuracy and the robustness of econometric modeling. The combination of linear and logistic formulations, along with supervised and unsupervised dimensionality reduction, exemplifies how econometric intuition and machine learning flexibility can jointly expand the toolkit for empirical economics.

6 - Conclusion

This study aimed to evaluate the predictive performance of different econometric and machine-learning techniques in high-dimensional settings, using the loan approval dataset as an empirical testing ground. The results demonstrate that shrinkage-based models, especially Ridge and Lasso, consistently outperform PCA and OLS, both in terms of accuracy and robustness. PLS also exhibited excellent performance, achieving the highest F1-score and accuracy, confirming its effectiveness as a supervised dimensionality-reduction method that directly leverages the covariance between predictors and the response variable.

All models were estimated in a binary classification framework. Logistic specifications were used for OLS, Lasso, and PCA-based models, while Ridge and PLS were implemented in their linear forms due to the unavailability of logistic equivalents. Their continuous predictions were transformed into probabilities and classified using a 0.5 threshold, allowing direct comparability of predictive outcomes across approaches.

From an applied perspective, these results highlight the importance of regularization and supervised dimensionality reduction in financial and economic forecasting problems, where large numbers of correlated predictors are common and model interpretability remains crucial. While PCA can still be valuable as a preprocessing or exploratory tool, the evidence suggests that shrinkage and PLS methods provide a better compromise between parsimony, interpretability, and predictive power.

In sum, the analysis reinforces the idea that modern econometric modeling benefits from the integration of machine-learning concepts, such as penalization, supervised factor extraction, and cross-validation, which improve traditional methods without sacrificing theoretical transparency. Together, Bayesian shrinkage and PLS-type approaches emerge as powerful, interpretable, and computationally efficient tools for economists and practitioners working with complex, data-rich environments.

7 - References

De Mol, C., Giannone, D., & Reichlin, L. (2008). Forecasting using a large number of predictors: Is Bayesian shrinkage a valid alternative to principal components? Journal of Econometrics, 146(2), 318–328. https://doi.org/10.1016/j.jeconom.2008.08.011

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12(1), 55–67. https://doi.org/10.1080/00401706.1970.10488634

Wold, H. (1975). Soft modelling by latent variables: The non-linear iterative partial least squares (NIPALS) approach. In Perspectives in Probability and Statistics (pp. 117–142). Academic Press.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer. https://doi.org/10.1007/978-0-387-84858-7

Stock, J. H., & Watson, M. W. (2002). Forecasting using principal components from a large number of predictors. Journal of the American Statistical Association, 97(460), 1167–1179. https://doi.org/10.1198/016214502388618960

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An introduction to statistical learning: With applications in R (2nd ed.). Springer. https://doi.org/10.1007/978-1-0716-1418-1

8 - Software and tools

The source code and dataset used in this project are publicly available on GitHub at the following repository:

GitHub Repository: *https://github.com/moiraclavin/ml-economists-final*

The repository contains:

- The main notebook: Machine_Learning_4_Eco.ipynb, which reproduces all the results and figures presented in this paper.
- The Loan Approval dataset (.csv) used for model training and validation.
- Auxiliary scripts for model evaluation (confusion matrices, ROC curves, and performance comparison).

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., … Oliphant, T. E. (2020). Array programming with NumPy. Nature, 585(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830. http://jmlr.org/papers/v12/pedregosa11a.html

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. Computing in Science & Engineering, 9(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

McKinney, W. (2010). Data structures for statistical computing in Python. In Proceedings of the 9th Python in Science Conference (pp. 51–56). https://doi.org/10.25080/Majora-92bf1922-00a

Van Rossum, G., & Drake, F. L. (2009). Python 3 reference manual. CreateSpace.

8 - Supplementary Material

Figure description:

1. Figure 1. Model performance comparison across methods.

Table summarizing the predictive performance of the five estimated models, OLS, PCA, Lasso, Ridge, and PLS, using classification metrics. The results indicate that all regularized and supervised dimensionality reduction techniques outperform PCA in terms of accuracy and F1-score, with PLS achieving the highest overall predictive performance. OLS serves as the unregularized benchmark, while Ridge and Lasso exhibit similar results, confirming the effectiveness of shrinkage in improving model stability and generalization.

2. Figure 2. Confusion matrices for the five estimated models.

Each panel displays the confusion matrix corresponding to the best specification of OLS, PCA, Lasso, Ridge, and PLS models. True negatives and true positives are shown along the main diagonal, while false positives and false negatives appear off-diagonal. The results indicate that PLS and Lasso achieved the highest classification accuracy, with fewer misclassifications compared to OLS and Ridge. PCA, in contrast, shows a higher number of false predictions, confirming its lower overall performance observed in the quantitative metrics.