Student Declaration of Authorship

HERIOT WATT UNIVERSITY
UK | DUBAI | MALAYSIA

Course code and name:	F21DL Data Mining and Machine Learning
Type of assessment:	**Group**
Coursework Title:	Machine Learning Portfolio
Student Name:	Nakhul Kalaivanan,Goktug Peker,Pietro Taliento
Student ID Number:	H00354508,H00479709,H00459440

Declaration of authorship. By signing this form:

- **I declare** that the work I have submitted for individual assessment OR the work I have contributed to a group assessment, is entirely my own. I have NOT taken the ideas, writings or inventions of another person and used these as if they were my own. My submission or my contribution to a group submission is expressed in my own words. Any uses made within this work of the ideas, writings or inventions of others, or of any existing sources of information (books, journals, websites, etc.) are properly acknowledged and listed in the references and/or acknowledgements section.

- I confirm that I have read, understood and followed the University's Regulations on plagiarism as published on the University's website, and that I am aware of the penalties that I will face should I not adhere to the University Regulations.

- I confirm that I have read, understood and avoided the different types of plagiarism explained in the University guidance on Academic Integrity and Plagiarism

Student Signature *(type your name): Nakhul,Goktug,Pietro*

Date: *21/11/2024*

Copy this page and insert it into your coursework file in front of your title page.
For group assessment each group member must sign a separate form and all forms must be included with the group submission.

Your work will not be marked if a signed copy of this form is not included with your submission.

F21DL Data Mining & Machine Learning Coursework Summary Report (Part 1)

Group_PG 14

Link to GitHub & Google Collab:

https://github.com/pietrotaliento/MLandDataMiningGroup14

https://colab.research.google.com/drive/1b8K2qNY5xEfDbEU-vMotYPsvlqYO23Q2#scrollTo=WlD114IWhjWS

H00354508	Nakhul Kalaivanan
H00459440	Pietro Taliento
H00479709	Goktug Peker

Cover Page – 1 page

Table of Contents – 1 page

Report – 6 page

Contribution – 1 page

Table of Contents

R1: Introduction - Project Topic & Directions

The predictive power of machine learning in financial analysis and image recognition has revolutionized how data-driven decisions are made. This report explores the application of advanced machine learning techniques across two diverse yet impactful datasets.

Dataset 1: Give Me Some Credit

The first dataset focuses on financial distress prediction, a critical concern for financial institutions aiming to assess the creditworthiness of individuals. By analysing demographic and financial attributes, this dataset enables the prediction of whether an individual is likely to experience serious financial delinquency within two years. Such predictions are invaluable for lenders in minimizing risks and tailoring financial products. The primary research questions for this dataset are:

1. **How does monthly income affect financial distress?**
2. **Can clustering techniques such as K-Means be applied to segment individuals?**

Using this dataset, we implemented a range of supervised and unsupervised machine learning techniques, including data preprocessing, clustering, and classification models, to derive actionable insights. Special attention was given to handling class imbalances, feature scaling, and model generalization to ensure robust predictions.

Dataset 2: Sign Language MNIST

The second dataset, Sign Language MNIST, addresses a different domain—static hand gesture recognition for the American Sign Language (ASL) alphabet. By leveraging image classification techniques, the dataset aims to bridge the gap in communication for individuals who rely on sign language. The research questions for this dataset are:

1. **How can sign language recognition models generalize across different individuals with varying hand shapes, speeds, and styles?**
2. **Can a unified machine learning model be trained to recognize signs from multiple sign languages and differentiate between them?**
3. **How can machine learning algorithms be developed to automatically recognize and translate sign language gestures into text?**

To address these questions, we experimented with both basic and advanced neural network architectures, including Multilayer Perceptrons (MLPs) and Convolutional Neural Networks (CNNs). These models were optimized for accuracy and robustness using techniques such as data augmentation, dropout, and hyperparameter tuning.

R2: Data Analysis

Data Analysis, Cleaning

After common imports, we loaded the dataset cs-training into our Google Collab common file.

We found 150000 instances as expected, divided in two classes based on if they did delinquency in the last two years or not.

Dataset has 11 different features including index column with unique numbers and 1 column for outcome (label) that makes our data set suitable for **classification problem for supervised learning.**

When we look at data points, we have observed null values in “Monthly income” column where data cleaning work is required.

Negative correlation was obtained as -21% between “age” and “number of depended” and highest correlation was obtained as 15% between ‘NumberofOpenCreditLineAndLoans’ and ‘age’.However as we understood the each features are generally independent and no specific high correlation was observed.

Data Cleaning and Handling Missing Values

We have a couple of options to deal with the *missing values*:

1.Drop the rows with incomplete values

2.Drop the attribute with missing values

3.Mean Imputation

4.Median Imputation

We learned about missing values in the 'NumberOfDependents' column and mostly in the 'MonthlyIncome' column, which needed to be handled. Since this attributes are crucial in the analysis, we decide to impute both with a mean, in order to maintain the integrity of the data.

Data Cleaning Conclusions

After careful consideration, we chose median imputation to handle missing values, as the missing data was concentrated in only two key features that were essential to our analysis. Given the presence of outliers and the importance of these features, we calculated the median for each affected column. This approach allowed us to retain as much data as possible while minimizing the influence of outliers. We then removed any rows with remaining missing values and re-examined the data through updated visualizations to ensure everything was in order. This strategy resulted in a complete and more consistent Data Frame, providing a solid foundation for generating meaningful visualizations and drawing valuable insights.

Final Notes: Conclusions on Data Analysis (EOD)

From our data analysis, we found that we are dealing with **a binary and fairly complex** dataset. The main challenges we addressed, and partially dealt with, include

- Clear Imbalance Between the Two Classes: Most individuals belong to class 0. We plan to address this imbalance using SMOTE
- Disparity in Feature Magnitudes: We will resolve this issue by applying scaling techniques
- Skewness in Certain Features: Features like 'RevolvingUtilizationOfUnsecuredLines' and 'DebtRatio' were skewed, but we handled this by applying log transformation
- Presence of Outliers: Some outliers, possibly due to measurement errors or unique individuals, were clipped. However, we will need to remain cautious as we proceed. Notably, 'Monthly Income' and 'Debt Ratio' exhibit many outliers, and since these are important features, we must ensure they are properly handled

Alternative Dataset

Similar to our main data set we applied the similar methods as we use median values for each feature to replace the missing values. Using median skewed the distribution and caused less distortion with skewness or outliers, preserving the central tendency closer to the actual "middle" of the data.

R3 – CLUSTERING

In this section we did clustering analysis on a tabular financial dataset, focusing on evaluating and optimizing the K-Means algorithm for customer segmentation. Experiments included varying the number of clusters, using PCA for dimensionality reduction, and calculating metrics like the Elbow Method and Silhouette Score to determine the optimal number of clusters.

Steps were followed before we started experimenting Clustering Algorithms;

1. Data Preprocessing
2. Feature Selection and Scaling
3. Outlier Removal

Experimenting with Clustering Algorithms and Optimal Clusters

K-Means Clustering

- Elbow Method
- Silhouette Score

The results for the Elbow method and Silhouette Scores are very clear. The highest Silhouette Score is given with 4 clusters, and Elbow method shows a clear slowing down in it's reduction at the same value, which, until futher expirements, makes it the ideal choice for us.

Other Evaluation Metrics:

- In addition to silhouette scores, the 'Calinski-Harabasz' and Davies-Bouldin scores were calculated to further evaluate clustering effectiveness. The 'Calinski-Harabasz 'score rewards high inter-cluster variance, while the Davies-Bouldin score favours lower intra-cluster distances.
- K-Means Silhouette Score: 0.3583
- K-Means 'Calinski-Harabasz' Score: 131706.1919
- K-Means Davies-Bouldin Score: 0.8510
- Silhouette Score indicates a decent clustering, which might have some overlapping and could also be perfected
- The clusters are compact, as shown in the high 'Calinski-Harabasz' Score
- Given the fairly low value in the Davies-Bouldin score, we somehow alleviate the worries about clusters overlapping too much, but there might still be shared boundaries

Evaluation and Insights

Cluster Evaluation:

- The Elbow Method and Silhouette Score plots helped verify that 4 clusters were optimal, balancing interpretability and cohesion.
- Using PCA, the separation between clusters was visually confirmed, with well-defined centroids and distinct cluster boundaries.

Cluster Interpretation:

- Each cluster can be interpreted based on financial traits such as income and debt ratio, allowing targeted customer segmentation. For instance, one cluster might represent low-income, high-debt individuals, while another represents high-income, low-debt customers. These insights could guide strategic decisions, such as targeted financial products for each group.

Final Notes

K-means, when combined with dimensionality reduction using PCA, can efficiently cluster high-dimensional data, as demonstrated by the well-defined, convex partitions in the PCA space. A potential limitation of K-means is its sensitivity to outliers, which can skew the centroids. However, it remains suitable for this dataset, as features like Debt Ratio and Monthly Income form dense regions of data

GMM is another excellent choice for this dataset, as it effectively handles the overlap between clusters, which was a primary challenge. Its probabilistic nature allows it to naturally capture relationships and the density of features, as well as address their non-linearities

Agglomerative Clustering, on the other hand, produced irregularly shaped clusters with poor separation. This method was less effective for a dataset with large dimensions and high feature complexity.

R4– BASIC CLASSIFIERS (DECISION TREE, RANDOM FOREST, XGBOOST)

Decision Tree Classifier

The initial model used a pruned Decision Tree Classifier to predict loan delinquency (SeriousDlqin2yrs). Although the model achieved interpretability and reduced overfitting, its performance was limited, particularly in handling imbalanced classes.

Step 1: Initial Decision Tree Model

Step 2: Improving Class Imbalance Handling Using SMOTE

Step 3: Model Stability Validation Using 10-Fold Cross-Validation

Step 4: Hyperparameter Tuning for Max Depth

Step 5: Expanding Testing Dataset for Better Generalization

Step 6: Experimenting with Different Train-Test Splits

Step 7: Comparative Analysis Using Random Forest

Step 8: Alternative Decision Tree Classifier approach, to increase the accuracy by avoiding overfitting using prune method

Final Notes

- The baseline Decision Tree model offered simplicity but struggled with class imbalance and generalization.
- SMOTE significantly improved minority class detection, as reflected in increased precision, recall, and F1-scores.
- Cross-validation ensured model stability and mitigated overfitting risks.
- **Random Forest emerged as the better-performing model compared to the Decision Tree, leveraging ensemble learning to improve robustness and accuracy.**

Random Forest Classifier

Enhancing Model Performance for Loan Delinquency Prediction

Stage 1: Addressing Overfitting and Underfitting

Stage 2: Evaluating Model Discrimination with ROC Curve

Stage 3: Gini Index vs. Entropy

Stage 4: Random Forest Feature Importance and Hyperparameter Tuning

Stage 5: Incorporating XGBoost for Performance Comparison

Final Notes

In the end, we successfully developed a reliable model to predict the target class with minimal or no overfitting. Given the imbalanced nature of our dataset, Random Forest proved to be an effective model in balancing the contributions of the classes and reducing overfitting, demonstrating its reliability for classification tasks in complex datasets like ours. SMOTE also showed to be a highly interesting and effective tool for that purpose, delivering immediate results.

We conducted several experiments on pruning the tree to find the optimal configuration. Metrics such as Precision, Recall, and F1-score provided a robust evaluation of the model's performance. As observed during data exploration, the NumberOfTimes90DaysLate feature emerged as the strongest predictor for the target class, though the contribution of other features like 'DebtRatio' should not be overlooked. Overall, we were satisfied with the results of our analysis.

R5–NEURAL NETWORK (MLP & CNN)

Multi-Layer Perception (MLP)

Objectives

The primary goal was to classify hand signs from the Sign Language MNIST dataset using a Multilayer Perceptron (MLP) model. While MLPs are not ideal for image data, the purpose was twofold:

1. Explore the feasibility of MLPs in this context.
2. Gain insights into data preprocessing and optimization techniques to inform future experiments with Convolutional Neural Networks (CNNs).

Final Notes

While the MLP was not the ideal choice for this task, it provided a solid starting point and valuable lessons for future work with CNNs. The outcomes highlight the importance of balancing simplicity and effectiveness, even when exploring less-than-ideal approaches.

Convolutional Neural Network (CNN)

Objectives

The goal was to classify hand signs from the Sign Language MNIST dataset using Convolutional Neural Networks (CNNs). This involved designing, training, and optimizing the CNN architecture to achieve high accuracy while minimizing overfitting. A secondary aim was to explore the impact of various hyperparameter configurations on the model's performance.

Methods

1. Data Preparation:
2. Data Augmentation:
3. Model Architecture:
4. Training and Validation:
5. Hyperparameter Optimization:

Final Notes

The CNN-based approach significantly outperformed MLPs and demonstrated the value of structured experimentation in achieving robust results. Overall, the project achieved its goals, providing a strong baseline for future image classification tasks while exploring the impact of various techniques on performance.

In the end, we achieved an optimal model with high accuracy on both the training and testing datasets. While the training results were slightly higher for smaller convolutional and hidden layer sizes, the overall performance remained consistent. The ReLU activation function proved to be the best choice for the CNN, as is commonly recognized.

Overall, we are highly satisfied with the final results and the experiments conducted to achieve them, as they allowed us to explore various techniques effectively.

REPORT CONCLUSION

Dataset 1: Give Me Some Credit

How does monthly income affect financial distress?

Analysis revealed that **Monthly Income** is a key determinant of financial distress, with individuals earning lower incomes being more likely to experience delinquency. Techniques such as **median imputation** preserved the integrity

of this feature despite missing values. Furthermore, feature importance analysis in **Random Forest** and **XGBoost** confirmed Monthly Income as one of the top predictors of financial risk.

Can clustering techniques such as K-Means be applied to segment individuals?

Clustering with **K-Means** proved effective in segmenting individuals based on their financial attributes. Using the **Elbow Method** and **Silhouette Scores**, we identified four optimal clusters representing distinct financial profiles, such as low-income, high-debt individuals and high-income, low-debt customers.

Dataset 2: Sign Language MNIST

How can sign language recognition models generalize across different individuals with varying hand shapes, speeds, and styles?

Generalization was achieved by employing **data augmentation** techniques, such as rotations, flips, and zooms, to simulate real-world variations in hand gestures. **Convolutional Neural Networks (CNNs)** demonstrated superior performance in recognizing diverse hand shapes and styles, achieving high accuracy with minimal overfitting. This approach ensures robustness across different user profiles, paving the way for broader adoption.

Can a unified machine learning model be trained to recognize signs from multiple sign languages and differentiate between them?

While this dataset focused on American Sign Language (ASL), the results suggest that **CNNs**, combined with **transfer learning** from pre-trained models, could be extended to recognize signs from multiple languages

How can machine learning algorithms be developed to automatically recognize and translate sign language gestures into text?

The **CNN-based model**, paired with a softmax output for classification, forms the foundation for real-time gesture-to-text translation. With additional training on dynamic gestures and integration into mobile or wearable devices, this solution could enable seamless translation of sign language into text for communication.

Best Solutions

For Dataset 1 (Financial Distress Prediction):

The **Random Forest** and **XGBoost models**, optimized with **SMOTE** to address class imbalance, emerged as the best classifiers. These models achieved high accuracy, precision, and AUC-ROC scores, effectively predicting delinquency while handling imbalanced datasets. Feature importance analysis provided actionable insights, identifying **NumberOfTimes90DaysLate** and **DebtRatio** as other key predictors alongside **Monthly Income**.

For Dataset 2 (Sign Language Recognition):

The **CNN architecture**, with **data augmentation** and **ReLU activation**, delivered the best results for sign language classification. This model achieved high accuracy on both training and test sets while maintaining robustness against overfitting.

CONTRIBUTIONS

#	Assignment Details	Nakhul	Pietro	Goktug	REMARK
R1	**Project Topic and Direction**				
R1.1	Chosing Data Set	☑	☑	☑	
R1.2	Preparing Presentation	☑	☑	☑	
R2	**Data Analysis**				
R2.1	Perform statistical analysis and visualizations to explore data patterns and relationships	☑	☑	☑	Code Running
R3	**Clustering**				
R3.1	Apply clustering techniques (e.g., K-Means) to group similar data points based on common features	☑	☑	☑	Group Study
R4	**Basic Classifiers and decission trees**				
R4.1	Implement basic classifiers and decision trees to classify data and visualize decision-making processes	☑	☑	☑	Group Study
R5	**Neural Networks (&CNN)**				
R5.1	Develop and apply neural networks, including convolutional neural networks (CNN), for data classification tasks, especially for images.	☑	☑	☑	Group Study