Danmarks Tekniske Universitet

DTU

02450 - Project 2

Authors

Nikolaos Iliakis - s250201
Pedro Francisco Martínez Bulacio - s243972
Álvaro Quintana López - s250202

April 10, 2025

0. Credit Attribution

0.1 Table

Section	**s250201**	**s243972**	**s250202**
1. Regression a	40%	30%	30%
2. Regression b	30%	30%	40%
3. Classification	30%	40%	30%
4. Discussion	40%	30%	30%

Table 1: Credit Attribution Table

Contents

1 Regression a

In this regression task, our objective is to predict the total amount spent by a customer (`MntTotal`), as this prediction can help the business better understand customer value and support targeted marketing strategies, using a set of continuous and categorical features that reflect both demographic and behavioral attributes of the customer–. In particular, we found it relevant to include regressors such as `Income`, `Age`, or the amounts spent on primary product categories like `MntFruits`, `MntMeatProducts` and `MntFishProducts`; among others, as these variables are likely to reflect both the customer's purchasing power and their consumption preferences. The selected input regressors for the model are:

- Continuous Variables: `Income`, `Recency`, `Age`, `Total_Children`, `Members_Household`, `Days_Enrolment`, `MntFruits`, `MntMeatProducts` and `MntFishProducts`.
- Categorical Variables: `Education` and `Living_Status`. The Education variable was one-hot encoded into three binary features: `Education_Graduate`, `Education_Postgraduate` and `Education_Undergraduate`. Similarly, the `Living_Status` variable was one-hot encoded into a binary variable `Living_Status_Alone`.

To ensure comparability among features and improve model performance, we applied a standardization transformation to the entire data matrix X, so that each column has zero mean and unit standard deviation. This transformation is particularly important as we will subsequently introduce regularization, which assumes that features are on the same scale.

To evaluate the generalization performance of the regularized linear regression model, we applied 10-fold cross-validation across a wide range of values $[10^{-1}, 10^{3}]$ for the regularization parameter λ. Each model was trained using standardized input features, followed by ridge regression with varying regularization strengths. The generalization error was estimated for each λ using the mean squared error (MSE) across the validation folds, and the model with the lowest error was selected as optimal.

However, the generalization error was reported in terms of root mean squared error (RMSE). We decided to use RMSE as the generalization error measure because it has the same units as the target variable `MntTotal` ($), making it easier to interpret in the context of the problem.

As observed in Figure 1, the plot of generalization error versus λ showed the typical bias-variance tradeoff, with the error decreasing initially as regularization helped prevent overfitting, and increasing again when the model became too constrained. The best model is obtained with a regularization parameter of $\lambda^* = 247$, resulting in a generalization error of 283 $ measured by the RMSE.

1.1 Linear Regression Model

The final linear regression model predicts the total amount spent by a customer (y) as a weighted sum of standardized customer attributes, as shown in Table 2.

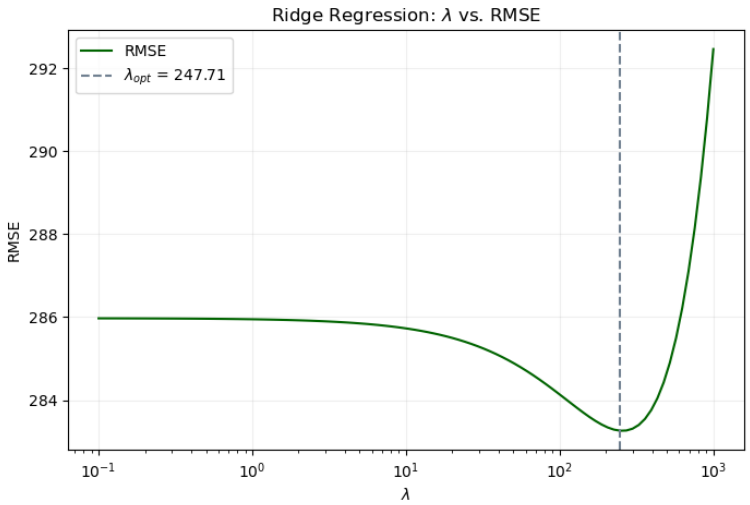

Figure 1: Generalization error as a function of regularization parameter λ for ridge regression

Each coefficient reflects the expected change in the predicted amount spent associated with an increase in the corresponding variable, holding all other variables constant.

Now, let's briefly analyze the effect of some of the individual attributes coefficients and their influence in the final model prediction.

Strong Positive Influences

- `Income` (129.08): The most influential feature. Higher income is strongly associated with increased spending.
- `Recency` (55.72): Indicates that long-term customers tend to spend more, possibly due to greater brand loyalty.
- `MntMeatProducts` (271.80): Customers who spend more on meat products also tend to spend more overall.

Strong Negative Influences

- `Total_Children` (-35.96): More children in a household are associated with reduced spending, likely due to tighter budget constraints.
- `Members_Household` (-23.50): Larger household sizes correlate with lower spending, possibly due to resource sharing or conservative consumption.
- `Education_Undergraduate` (-22.83): Suggests that individuals with undergraduate education tend to spend less compared to other education levels.

Overall, the ridge regression model captures several intuitive patterns. Income and food-related purchases are strong predictors of spending. Longer enrollment and age also correlate positively, suggesting loyal or older customers tend to spend more. In contrast, indicators of financial strain

Feature	**Coefficient**
Income	129.08
Recency	6.21
Age	29.10
Total_Children	-35.96
Members_Household	-23.50
Days_Enrolment	55.72
Education_Graduate	-10.70
Education_Postgraduate	25.94
Education_Undergraduate	-22.83
Living_Status_Alone	-11.80
MntFruits	78.70
MntMeatProducts	271.80
MntFishProducts	92.85

Table 2: Ridge Regression Coefficients

or shared consumption (like children or household size) tend to reduce predicted spending, as well as the education. Furthermore, some variables (education-related) are weakly correlated with the target, which can limit the predictive power of the model. We will observe this limitation more clearly in Exercise 3: Classification. Regularization helps reduce overfitting, but the model's generalization ability might still be constrained by the quality and size of the data.

2 Regression b

To evaluate whether more complex models outperform the ridge regression model developed in Part A, we implemented a two-level cross-validation (CV) procedure to compare three models:

- A regularized linear regression model (Ridge)
- An artificial neural network (ANN)
- A baseline model that simply predicts the mean amount spent

This comparison aims to answer two key questions:

1. Is there a model that consistently outperforms the others?
2. Are the Ridge and ANN models significantly better than a trivial baseline?

We applied **two-level cross-validation** with $K_1 = K_2 = 10$, following the structure in Algorithm 6 from the lecture notes. In each outer fold:

- The training set D_i^{par} is used for model training and **inner CV** hyperparameter tuning.
- The best models are evaluated on D_i^{test} to estimate generalization performance.
- All three models (Ridge, ANN, Baseline) use the **same outer folds**, ensuring fair and paired comparisons.

The evaluation metric used is the **mean squared error (MSE)** per observation, reported as **RMSE** to match the units of the target variable $y = \text{MntTotal}$.

2.1 Model Implementations and Hyperparameter Selection

Ridge Regression

The regularized linear model was tuned using `GridSearchCV` over a range of $\lambda \in [10^{-1}, 10^4]$ (20 log-spaced values). As in Part A, the model uses standardized features, and we reused the exact folds to ensure comparability.

ANN (Artificial Neural Network)

The ANN was implemented using PyTorch [2]. It consists of:

- One hidden layer with variable size h,
- ReLU activation,
- An output layer with one neuron (for regression),
- Trained with the **Adam optimizer** and MSE loss.

The complexity of the network is controlled via the number of hidden units h, which was tuned over the set {10, 20, 25, 30}. Regularization was applied using `weight_decay=1e-4`. Each ANN was trained for a fixed number of 5000 iterations. Early stopping was not used.

Baseline Model

Implemented using `DummyRegressor`, it simply predicts the mean of the training set y for all test samples.

2.2 Results

The table below shows the outcome of the 10 outer folds. For each fold i, we report:

- The optimal hidden layer size h_i^* selected for the ANN,
- The optimal regularization parameter λ_i^* selected for Ridge,
- The corresponding RMSEs on the outer test sets for all three models.

Outer i	ANN		Linear regression		Baseline
	h_i^*	E_i^{test}	λ_i^*	E_i^{test}	E_i^{test}
1	30	292.50	263.67	268.86	603.97
2	25	286.98	263.67	271.34	647.31
3	25	273.09	263.67	250.52	628.22
4	30	627.87	42.81	470.30	586.09
5	30	242.97	263.67	219.73	588.06
6	20	292.64	263.67	248.70	614.20
7	30	284.40	263.67	252.94	608.79
8	25	278.91	263.67	251.20	572.28
9	30	284.63	263.67	264.27	579.97
10	30	299.81	143.84	270.74	591.53
Average	**27.5**	**317.20**	**209.11**	**276.36**	**602.64**

Table 3: Outer fold (i) estimated test errors E_i^{test} and selected hyperparameters for each model using 10-fold nested CV. For ANN, h^* represents the optimal number of hidden units, and for Linear regression, λ^* represents the optimal regularization parameter.

Although the exact value of λ^* differs slightly, the results are highly consistent with those from the previous section. In most outer folds, the inner cross-validation selected $\lambda^* = 263.67$, which is very close to the previously found value of $\lambda^* = 247$. This small difference can be attributed to variability across folds, but overall, it confirms that the model's preferred regularization strength is **robust across different validation setups**.

2.3 Statistical Evaluation of Models

To validate the model comparison results statistically, we applied **paired t-tests** (Setup I), comparing test errors across the 10 outer folds. Since each model was evaluated using the **same**

outer splits, we can treat the test RMSE values as paired samples and test whether one model significantly outperforms another.
We conducted the following pairwise comparisons:

Comparison	Mean Difference	t-statistic	p-value	95% Confidence Interval	Conclusion
ANN vs Ridge	-39.52	-2.96	0.0159	[−69.69, −9.35]	Ridge significantly better
ANN vs Baseline	+285.66	+7.67	< 0.0001	[+201.43, +369.90]	ANN significantly better
Ridge vs Baseline	+325.18	+13.30	< 0.0001	[+269.86, +380.51]	Ridge significantly better

These results lead to the following interpretations:

- The **difference between Ridge and ANN is statistically significant** ($p = 0.0159$). The confidence interval for the difference does not include 0 and is entirely negative, indicating that **Ridge Regression consistently outperforms the ANN** on this dataset.
- Both **ANN and Ridge are significantly better than the baseline** ($p \ll 0.01$), with large positive differences in RMSE. This confirms that **both models learn meaningful structure** in the data, as opposed to simply predicting the mean.

These findings provide statistical confirmation of the trends observed in the raw cross-validation results. The t-tests reinforce the conclusion that Ridge Regression is the **most reliable and accurate model** for this problem. he ANN model remains a viable alternative, particularly in settings with larger datasets or nonlinear patterns, but in this case, its higher variance and training sensitivity limited its effectiveness. The baseline model serves only as a point of reference and should not be used for prediction.

3 Classification

3.1 Initial Approach

During the initial phase of our modeling process, we aimed to classify customers based on their education level (Undergraduate, Graduate, Postgraduate) using a selection of numerical and categorical features. However, after performing two-layer cross-validation on logistic regression and classification tree models, we observed that the models consistently performed at a level very similar to the baseline classifier. This indicated that the models were unable to learn patterns beyond what could be achieved through basic guessing (baseline).
To understand the root cause of this behavior, we analyzed the correlation matrix (computed in Project 1) between the available features and the education levels (Undergraduate, Graduate and Postgraduate). The results showed that there was no strong correlation between the target variable (Education) and any of the predictor variables. Most correlations with the education categories ranged between -0.2 and 0.1, suggesting a lack of meaningful linear relationships. This helped explain the models' inability to outperform the baseline (they were not provided with features that carried a sufficient predictive signal with respect to the target).

3.2 Classification Task: Predicting Family Size

In this classification task, our objective is to predict the `FamilySize` of a customer, categorized as either `Large` or `Small`, based on a combination of behavioral and demographic features. This is a binary classification problem. A family is considered `Large` if the family members are more or equal to 3 and `Small` if they are less than 3.
The input features selected for the model are:

- Continuous Variables
  `Income`, `Recency`, `Age`, `Days_Enrolment`, `MntFruits`, `MntMeatProducts`, `MntFishProducts`, `MntTotal`
- Categorical Variables (One-Hot Encoded)
  `Education_Graduate`, `Education_Postgraduate`, `Education_Undergraduate`

All continuous variables were standardized to have zero mean and unit variance. Categorical variables were one-hot encoded. This preprocessing ensures comparability and enhances the performance of models with regularization.
To address this binary classification problem, we compared the performance of three models:

- **Logistic Regression**: Regularization strength λ served as the complexity-controlling parameter, with values chosen on a log scale in the range $[10^{-3}, 10^{3}]$.
- **Baseline Model**: This model predicts the most frequent class from the training data and serves as a lower-bound reference for comparison.
- **Classification Trees (Method 2)**: Tree depth was used as the complexity-controlling parameter, and a range of depths from 5 to 20 was explored.

3.3 Model Evaluation: Two-Level Cross-Validation

To estimate generalization performance and conduct model selection using scikit [3], we applied 2-layer cross-validation with $K_1 = K_2 = 10$ folds. Each outer fold was used to estimate test error, while the inner loop was used to optimize hyperparameters:

- Tree depth d for Classification Trees
- Regularization strength λ for binary logistic regression

The results of the cross-validation procedure are summarized in Table 3. Each row represents an outer fold, showing the selected hyperparameters and corresponding test errors.

Outer i	Classification Trees		Logistic Regression		Baseline
	d_i^*	E_i^{test}	λ_i^*	E_i^{test}	E_i^{test}
1	8	0.3108	0.0010	0.3198	0.4595
2	11	0.2387	0.0010	0.2748	0.4505
3	6	0.2658	0.0016	0.2748	0.4550
4	5	0.3167	0.0026	0.2851	0.4299
5	7	0.2986	0.0016	0.3032	0.4208
6	7	0.2986	0.0010	0.3348	0.4842
7	9	0.2941	0.0016	0.2670	0.4615
8	8	0.3032	0.0026	0.3348	0.4434
9	13	0.3032	0.0010	0.3167	0.4389
10	9	0.3077	0.0010	0.3439	0.5023
Average	**8**	**0.2932**	**0.0015**	**0.3057**	**0.4547**

Table 4: Outer fold (i) estimated test errors E_i^{test} and selected hyperparameters for each model using 10-fold 2-layer CV. For Classification Trees, d^* represents the optimal depth, and for Logistic Regression, λ^* represents the optimal regularization parameter.

3.4 Statistical Evaluation of Classification Models

To assess whether the observed differences in performance between the models are statistically significant, we performed pairwise comparisons using **McNemar's test** (setup I) with a significance level of $\alpha = 0.05$. For each pair, we report the comparison matrix, the point estimate $\theta = \theta_A - \theta_B$, the 95% confidence interval and the p-value from the exact binomial test. The results are shown in Table 5.

Overall Conclusion: Both **logistic regression** and **classification trees** significantly outperform the **baseline** model. However, there is **no significant difference** in performance between logistic regression and classification trees. Therefore, either model is suitable for deployment: logistic regression is advantageous for its interpretability, while classification trees provide intuitive decision structures and visual insights.

	Baseline vs Logistic	Baseline vs CT	Logistic vs CT
Comparison Matrix	$\begin{bmatrix} 91 & 19 \\ 54 & 57 \end{bmatrix}$	$\begin{bmatrix} 96 & 14 \\ 57 & 54 \end{bmatrix}$	$\begin{bmatrix} 123 & 22 \\ 30 & 46 \end{bmatrix}$
Point Estimate θ	-0.1674	-0.1945	-0.0271
CI ($\alpha = 0.05$)	$[-0.2306,\ -0.0853]$	$[-0.2641,\ -0.1241]$	$[-0.0997,\ 0.0275]$
p-value	$5.06 \cdot 10^{-5}$	$2.67 \cdot 10^{-7}$	0.332

Table 5: Statistical comparison of classification models using McNemar's test.

3.5 Final Logistic Regression Model

We trained a logistic regression model to classify whether a customer belongs to a `Small` or `Large` family using the optimal value of $\lambda = 0.001610$ obtained from nested cross-validation (see Section 4). The corresponding value of $C = 1/\lambda$ was used in the model to control regularization strength and prevent overfitting.

How the Model Makes Predictions Logistic regression computes a weighted linear combination of the input features:

$$z = \mathbf{w}^\top \mathbf{x} + b$$

This score z is then passed through the sigmoid function:

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

to obtain a probability between 0 and 1. The model assigns a class label based on a threshold (typically 0.5): if $\sigma(z) \geq 0.5$, the customer is classified as `Large`; otherwise, as `Small`.

Feature Relevance Feature relevance was assessed by examining the absolute value of the learned coefficients. The most influential feature was:

- `Education_Undergraduate`, which had the largest positive coefficient, indicating that customers with undergraduate education are more likely to belong to a `Large` family.

Other features such as `Income`, `MntTotal`, `Days_Enrolment`, and product-specific spending had near-zero coefficients, suggesting low predictive value in the classification context.

Comparison with the Regression Task In contrast to the regression task earlier in the report (predicting `MntTotal`), the most relevant features were `Income`, `Days_Enrolment`, and `MntMeatProducts`. These features were strong predictors of total spending but did not contribute significantly to predicting family size.
This contrast highlights that while some features (such as income and spending) are important for modeling financial behavior, the educational background, specifically being an undergraduate, is more informative for estimating family size.

4 Discussion

Regression The regression analysis provided valuable insight into the performance of Ridge regression, a simple artificial neural network (ANN), and a baseline model. The use of two-level cross-validation ensured a reliable comparison and allowed for the selection of optimal hyperparameters while minimizing overfitting.
Ridge regression proved to be the most consistent and reliable model, achieving the lowest generalization error across folds. The selected regularization values were stable and closely aligned with those found in Part A, reinforcing the model's robustness.
The ANN also outperformed the baseline but was slightly less effective than Ridge. This is likely due to the relatively simple ANN architecture used, a single hidden layer with a small number of units. With a deeper network or more tuning, the ANN could potentially capture more complex patterns and improve its performance.
Statistical tests confirmed that Ridge significantly outperformed the ANN, and both models were significantly better than the baseline. These results emphasize the importance of matching model complexity to data characteristics. Ridge offered a strong, interpretable solution, while the ANN remains a promising candidate for future improvement.

Classification Throughout this classification process, we gained several valuable insights about the modeling process. Initially, we attempted to classify education levels but found that the models could not outperform the baseline. This led us to reflect on the limitations of the task itself, understanding that no matter how sophisticated the model is, its predictive performance is limited by the quality and relevance of the input features. Consequently, we had to recognize the importance of re-evaluating our objectives when the initial results were unproductive.
By shifting our focus to predicting family size, we were able to construct a more meaningful classification task that both logistic regression and classification trees handled effectively. The fact that both models significantly outperformed the baseline, yet showed no statistically significant difference between them, highlighted that multiple modeling approaches can perform similarly well in practice. This opens the door to choosing models not only based on performance, but also based on interpretability or implementation.

We reviewed the original analysis and documentation of the Customer Personality Analysis dataset. The source material primarily focuses on descriptive statistics and exploratory data analysis, such as customer segmentation based on demographics and product preferences. However, it does not include any application of supervised learning methods such as classification or regression. As a result, we were unable to find any prior studies using this dataset for predictive modeling tasks. Therefore, the comparison with existing classification or regression studies is not applicable in this case.

The text in this document was refined with assistance from AI-based language models such as ChatGPT [1].

References

[1] OpenAI, "ChatGPT: Language Model for Text Generation," OpenAI, 2024. [Online]. Available: `https://openai.com/chatgpt`

[2] PyTorch, "PyTorch Documentation," 2024. [Online]. Available: `https://pytorch.org/docs/stable/`

[3] Scikit-learn Developers, "scikit-learn: Machine Learning in Python - User Guide," 2024. [Online]. Available: `https://scikit-learn.org/stable/user_guide.html`