

Assignment 1

Report

Name: Jonas Schrade
Student Number: 01/1080887
Course (Instructor): Deep Learning for Social Science (Giordano Di Marzo)

1 Introduction

The goal of assignment 1 is to build, train, and evaluate different multi-layer perceptron (MLP) models for predicting continuous salary data and discussing their performance with respect to a benchmark linear regression.

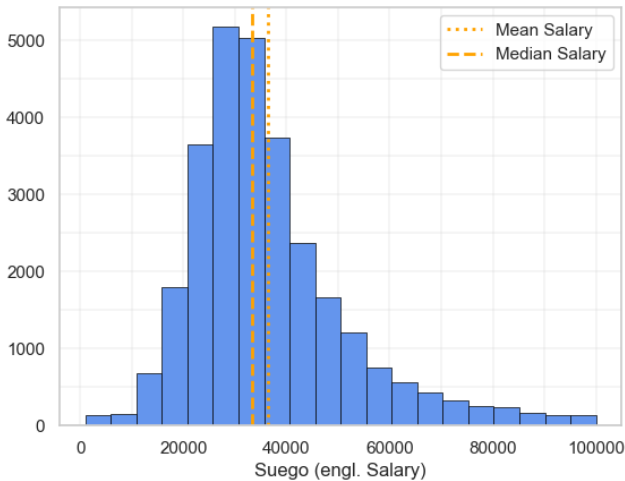


Figure 1: Target variable histogram.

2 Results

2.1 Data

The dataset used to train, validate, and test the models consists of 28,631 Uruguayan vacancies with 31 attributes and is provided by the course instructor who obtained it from the Internal Labour Organization. The target/predicted variable of the regression, the "Suego" (engl.salary) of the vacancy, has its mean at \$36,461.83 and a median of \$33,241.14, distributing over a large range of values (Figure 1). To gain a comprehen-

sive understanding of the remaining variables intended as regression features, an exploratory data analysis (EDA) is performed. The dataset comprises six continuous and six categorical variables, alongside sixteen binary indicators related to vacancy characteristics. These binary variables are further classified into four distinct categories: confidential, cogn, social, and manual. The job-specific features capture information on required skills, language proficiency, education level, and location. The distribution of all features is visualized in the accompanying assignment notebook. While correlations among the numerical features are generally weak (Figure 2), some variables exhibit substantial amounts of missing data (Figure 3).

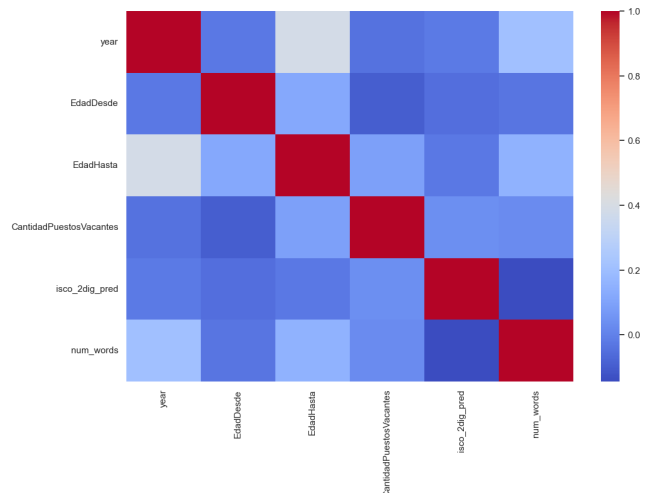


Figure 2: Correlation matrix (numeric).

To address data weaknesses caused by missing values, imputation is performed. Categorical variables are imputed with the mode, while continuous variables are imputed with

the median, as recommended by the instructor. Regarding the large range of the target variable and potential right-skewness, it is put into the natural logarithm. Additional preprocessing steps include hot-decking the categorical variables, separating the dataset into features and the regressand, and splitting it into training, validation, and test sets. A scaler is fitted to the training set, and continuous features are standardized consistently across the training, validation, and test sets. Finally, the processed datasets are stored as Torch tensor objects.

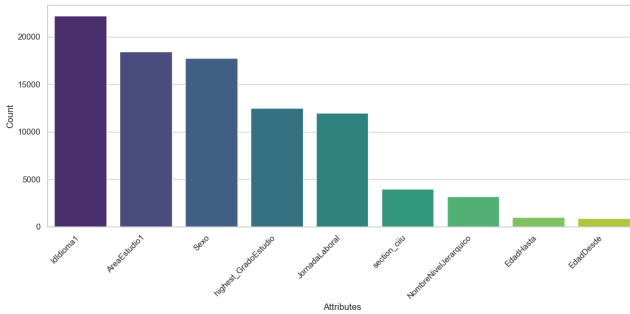


Figure 3: Variables with >500 missing values.

2.2 Models

To evaluate the performance of neural network architectures, three different multilayer perceptron (MLP) models are designed and compared against a linear regression benchmark. The linear regression model, trained on the same dataset, achieved a Mean Squared Error (MSE) of 0.142 and a Mean Absolute Error (MAE) of 0.254 when predicting salary values on the validation set.

Model 1 represents a conventional MLP architecture with four hidden layers, structured with 128, 64, 32, and 16 neurons, respectively. Each hidden layer utilizes the ReLU (Rectified Linear Unit) activation function, followed by a final linear output layer. The model is trained with a learning rate of 0.0005, a batch size of 128, and for 20 epochs.

Model 2 is designed to illustrate the principles of the Universal Approximation Theorem, which states that a single hidden layer with a sufficient number of neurons can approxi-

mate any continuous function. Accordingly, this model comprises a single hidden layer with 240 neurons activated by ReLU, followed by a linear output layer. Hyperparameters are set to match Model 1, with a learning rate of 0.0005, a batch size of 128, and 20 epochs.

Model 3 shares the same architecture as Model 1 but adopts a more aggressive training strategy. The learning rate is increased to 0.05, and the batch size is raised to 1718, with the number of epochs reduced to 10. This configuration is intended to speed up the training process while exploring the impact of faster gradient updates and larger batch generalizations.

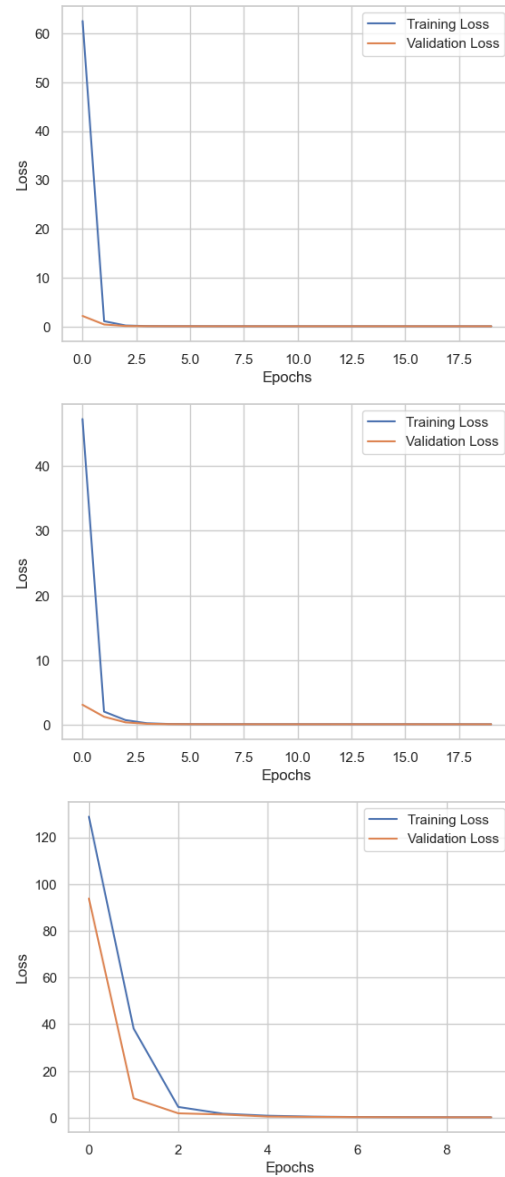


Figure 4: Regression task: Learning curves of Model 1, Model 2, Model 3 (top-bottom).

2.3 Training

For all three models, notable initial drops in the learning curves (for both training and validation loss) were observed (see Figure 2.2). In Model 1, the training loss decreases from 62.48 to 1.17 during the first epoch, eventually stabilizing at approximately 0.14 for both training and validation losses after 20 epochs. Model 2, on the other hand, experiences a rapid decrease in training loss from 47.2 to 2.1 in the first epoch, ultimately reaching a similar loss value after 20 epochs. Model 3 exhibits a slower reduction in training loss: initially dropping from 128.9 to 38.2 in the first epoch, and then to 4.6 by the second epoch. However, by the end of the 10 epochs, both training and validation losses converge to approximately 0.17, suggesting that with additional training, it could perform similarly to Model 1 and Model 2. Importantly, overfitting does not appear to be a concern for any of the models, as the validation loss remains low even with extended training durations.

2.4 Evaluation

The performance of the MLP regression models is evaluated on the validation set by comparing their Mean Absolute Error (MAE) and Mean Squared Error (MSE). Models 1 and 2 demonstrate similar outcomes, with Model 2—an MLP with a single hidden layer—achieving a slightly lower MSE of 0.138 compared to 0.139 for Model 1. However, when the training process is repeated, there is no clear performance advantage between the two models, as Model 1 occasionally records a lower error. Nevertheless, for the purpose of analysis, Model 2 is selected as the MLP with the best performance. In contrast, Model 3 records a higher MSE of 0.169, indicating weaker performance.

3 Conclusion

To evaluate the performance of the preferred MLP against the benchmark model, the MSE and MAE of both the MLP and linear regression are compared on the test set. In this

case, the MLP outperforms with an MSE of 0.136, slightly lower than the 0.14 observed for linear regression. However, this difference is marginal, raising questions about whether the increased complexity and computational cost of the MLP are justified compared to the simpler linear regression model.

4 Bonus

4.1 Random Forest Regressor

For further benchmarking, a random forest regressor is also trained on the data, serving as an additional machine learning alternative to the MLP. When applied to the test data for salary prediction, it achieves an MSE of 0.116, outperforming both the linear regression and the MLP models.

4.2 Ensemble Method

An ensemble of the three MLP models (as previously described) is constructed and evaluated against the performance of the individual MLPs. Using the validation data, the ensemble achieves an MSE of 0.14, which is slightly higher than the MSE of individual Models 1 and 2, but an improvement over the performance of Model 3. Additionally, the ensemble outperforms the average MSE of the individual models, which is 0.1486.

4.3 Classification

Alternatively, the MLP models are adapted to a classification task. For this, the target variable Salary is categorized into a binary indicator taking up the value of 1 if a vacancy offers more than \$30,000 and 0 if its less or equal. A logistic regression and random forest classifier are fit on the training data for benchmark. The general architecture of all three models remains the same, however, a sigmoid activation function is added to the output layer to produce binary classification output. Furthermore, instead of MSE, Binary Cross Entropy (BCE) will serve as loss criterion for the gradient descent.

Metric	Model 1	Model 2	Model 3
F1 Score	0.8168	0.8178	0.8121
Accuracy	0.7541	0.7511	0.7482
Precision $\leq 30k$	0.76	0.78	0.75
Recall $\leq 30k$	0.53	0.50	0.53
Precision $> 30k$	0.75	0.74	0.75
Recall $> 30k$	0.89	0.91	0.89

Table 1: MLP performance classification of validation set.

The learning curves of the three models differ significantly from those observed in the regression analysis (see Figure 5). For both Model 1 and Model 2, the validation loss plateaus around 0.5, with Model 1 displaying a slight upward trend, which may suggest potential overfitting. Model 3 exhibits a comparable pattern; however, its learning curve is arguably less informative due to the limited number of training epochs. Performance evaluation on the validation set reveals relatively consistent results across the models, with Model 2 achieving a marginally higher F1 score compared to the others (Table 1). When assessed against the benchmark, the outcomes mirror those of the regression task: the Random Forest model achieves the highest F1 score and overall accuracy, while the MLP model ranks second in classification performance (Table 2).

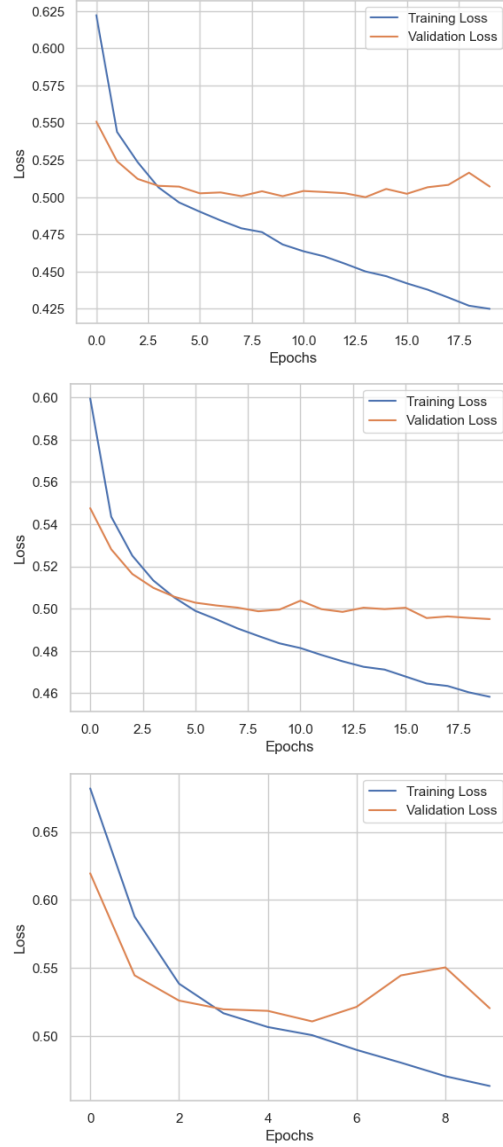


Figure 5: Classification task: Learning curves of Model 1, Model 2, Model 3 (top to bottom).

Metric	Logit	RF	Model 2
F1 Score	0.8030	0.8442	0.8244
Accuracy	0.7180	0.7866	0.7521

Table 2: Classification performance: Benchmark vs. MLP 2.