Assignment 2

Model Evaluation and Deployment

**Learning Outcome Addressed**

* Define the machine learning life cycle
* Explain feature engineering and the differences between nominal, ordinal and text data
* Analyse model evaluation with regard to training, validation and metrics
* Explain model deployment and database design

You have learnt about the machine learning life cycle and explored each stage in detail. Based on your learning, try and propose an ML approach to solve a hypothetical data science problem.

Let's assume you have to build a model to approve or reject loan requests. In this context,  answer the following:

**Questions**

* What type of data would you collect?
* How would you implement feature engineering?
* How would you deal with overfitting?
* How would you train and test your model?
* What metrics would you use?

**Submission Instructions**

* Download this assignment to record your responses.
* After you complete your assignment, select the **Start Assignment** button at the top of the assignment page.
* Upload the document containing your response.
* Select the **Submit Assignment** button to submit your responses

**Suggested time:** 120 minutes

***Note:****This is a required assignment and counts towards programme completion.*

**SMU AIML Required Assignment 6.3: Model Evaluation and Deployment**

| SMU AIML Required Assignment 6.3: Model Evaluation and Deployment | | |
| --- | --- | --- |
| **Criteria** | **Ratings** | **Pts** |
| Identify the type of data to collect | |  |  | | --- | --- | | **2 pts**  **Meets expectations**  Identified the type of data to be collected | **0 pts**  **Did not meet expectations**  Did not identify the type of data to be collected | | 2 pts |
| Explain the implementation of feature engineering | |  |  | | --- | --- | | **2 pts**  **Meets expectations**  Explained the implementation approach for feature engineering | **0 pts**  **Did not meet expectations**  Did not explain the implementation approach for feature engineering | | 2 pts |
| Analyse how to deal with overfitting | |  |  | | --- | --- | | **2 pts**  **Meets expectations**  Explained the plan to deal with overfitting | **0 pts**  **Did not meet expectations**  Did not explain the plan to deal with overfitting | | 2 pts |
| Explain how to train and test model | |  |  | | --- | --- | | **2 pts**  **Meets expectations**  Explained the approach to train and test the model | **0 pts**  **Did not meet expectations**  Did not explain the approach to train and test the model | | 2 pts |
| List the metrics to be used | |  |  | | --- | --- | | **2 pts**  **Meets expectations**  Listed the metrics to be used | **0 pts**  **Did not meet expectations**  Did not list the metrics to be used | | 2 pts |
| Total Points: 10 | | |

Required Assignment   
Model Evaluation and Deployment

**<Your Name>**

|  |  |
| --- | --- |
| Clock | Suggested time: 120 minutes |

**Task:** You have learnt about the machine learning life cycle and explored each stage in detail. Based on your learning, try and propose an ML approach to solve a hypothetical data science problem.

Let's assume you have to build a model to approve or reject loan requests. In this context, reflect upon and answer the following:

**Questions**

* What type of data would you collect?
* How would you implement feature engineering?
* How would you deal with overfitting?
* How would you train and test your model?
* What metrics would you use?

***Note:****This is a required assignment and counts towards programme completion.* *The solution to this assignment will be made available after you submit the assignment.*

|  |
| --- |
| 1. Type of Data to Collect To build a model for approving or rejecting loan requests, the following types of data should be collected:   * **Demographic Data:** Age, gender, marital status, number of dependents. * **Financial Data:** Income, employment status, loan amount requested, loan purpose, credit history, existing debts. * **Behavioral Data:** Payment history, transaction history, spending patterns. * **Loan Data:** Previous loan amounts, loan repayment history, loan type, interest rates.  2. Implementation of Feature Engineering Feature engineering involves creating new features or modifying existing ones to improve model performance. Steps include:   * **Handling Missing Values:** Using mean, median, or mode imputation, or more sophisticated methods like K-Nearest Neighbors (KNN) imputation. * **Encoding Categorical Data:** Using one-hot encoding for nominal data and ordinal encoding for ordinal data. * **Scaling Numerical Data:** Standardizing features by removing the mean and scaling to unit variance using StandardScaler or MinMaxScaler. * **Creating New Features:** Deriving new features like debt-to-income ratio, length of credit history, or interaction terms. * **Text Data Processing:** Converting text data into numerical representations using techniques like TF-IDF (Term Frequency-Inverse Document Frequency) or word embeddings.  3. Dealing with Overfitting Overfitting can be addressed using several techniques:   * **Regularization:** Applying L1 (Lasso) or L2 (Ridge) regularization to penalize large coefficients. * **Cross-Validation:** Using k-fold cross-validation to ensure the model generalizes well to unseen data. * **Pruning (for Tree-based Models):** Reducing the complexity of decision trees by limiting their depth or by pruning. * **Dropout (for Neural Networks):** Randomly dropping units (along with their connections) during training to prevent co-adaptation. * **Early Stopping:** Monitoring the model's performance on a validation set and stopping training when performance stops improving.  4. Training and Testing the Model The approach to training and testing the model includes:   * **Splitting the Data:** Dividing the dataset into training, validation, and test sets (e.g., 70% training, 15% validation, 15% test). * **Model Selection:** Choosing appropriate algorithms such as logistic regression, decision trees, random forests, gradient boosting, or neural networks. * **Hyperparameter Tuning:** Using grid search or randomized search with cross-validation to find the optimal hyperparameters. * **Training:** Fitting the chosen model to the training data. * **Validation:** Evaluating the model's performance on the validation set and tuning hyperparameters accordingly. * **Testing:** Assessing the final model performance on the test set to estimate its generalization capability.  5. Metrics to Use Key metrics for evaluating the loan approval model include:   * **Accuracy:** The proportion of correctly classified instances out of the total instances. * **Precision:** The proportion of true positive predictions among all positive predictions. * **Recall (Sensitivity):** The proportion of true positive predictions among all actual positives. * **F1 Score:** The harmonic mean of precision and recall, providing a balance between the two. * **ROC-AUC:** The area under the receiver operating characteristic curve, measuring the trade-off between true positive rate and false positive rate. |