MSIN0097: Predictive Analytics 24/25
MSIN0097 Individual Coursework
Word Count: 1875

Introduction

The dataset used in this project comes from a Portuguese bank's direct marketing campaign, where the objective was to promote the bank's term deposit product to potential customers. Since the marketing strategy involved telephone calls, the dataset includes various variables related to customer demographics and their interaction history with the campaign.

The key research question is: Can we predict whether a client will subscribe to the term deposit product based on their background information and campaign interaction data? This is a binary classification problem. In this case, the target variable (y) has two possible values: 1 (yes) if the client subscribes to the term deposit and 0 (no) if they do not.

The dataset consists of 41,188 records and includes multiple features that can be categorized into two main groups: (1) customer demographic information, such as age, job, marital status, education level, and housing loan status, and (2) campaign interaction data, such as the number of times the customer was contacted and the outcome of previous contacts. By using machine learning methods, we want to develop a predictive model that can classify customers into those who are likely to subscribe and those who are not.

In this project, I will explore different machine learning models and evaluate their performances using key metrics to identify the most effective approach for this classification problem.

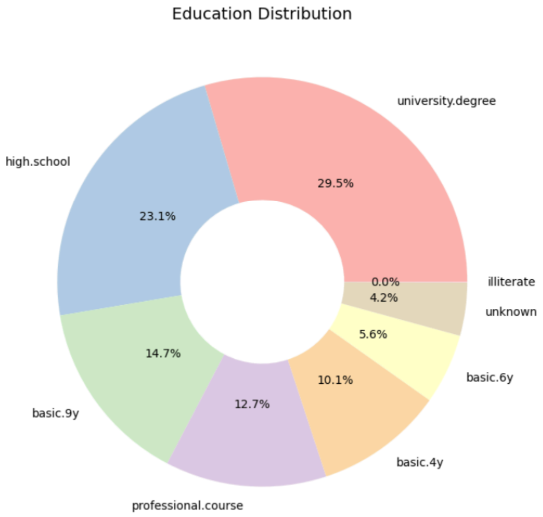
Data Exploration and Visualization

Before starting to build a predictive model, some graphs can help understand the dataset's characteristics through exploratory data analysis (EDA). This helps us identify trends and patterns in the data, which can later inform our model selection and feature engineering decisions.

To achieve this, I utilize three key visualizations:

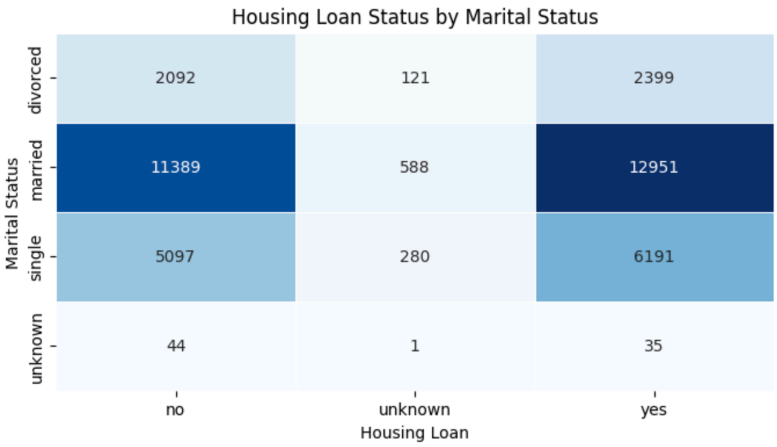
1. Education Distribution (Pie Chart)

The education distribution pie chart shows the proportion of customers with different educational backgrounds. As shown in the image below, university degree holders form the largest group (29.5%), followed by high school graduates (23.1%). Other education levels, such as basic education (4, 6, and 9 years) and professional courses, also make up a significant portion of the dataset. Remaining a small percentage of customers have unknown or illiterate educational status.



2. Heatmap of Housing Loan Status by Marital Status

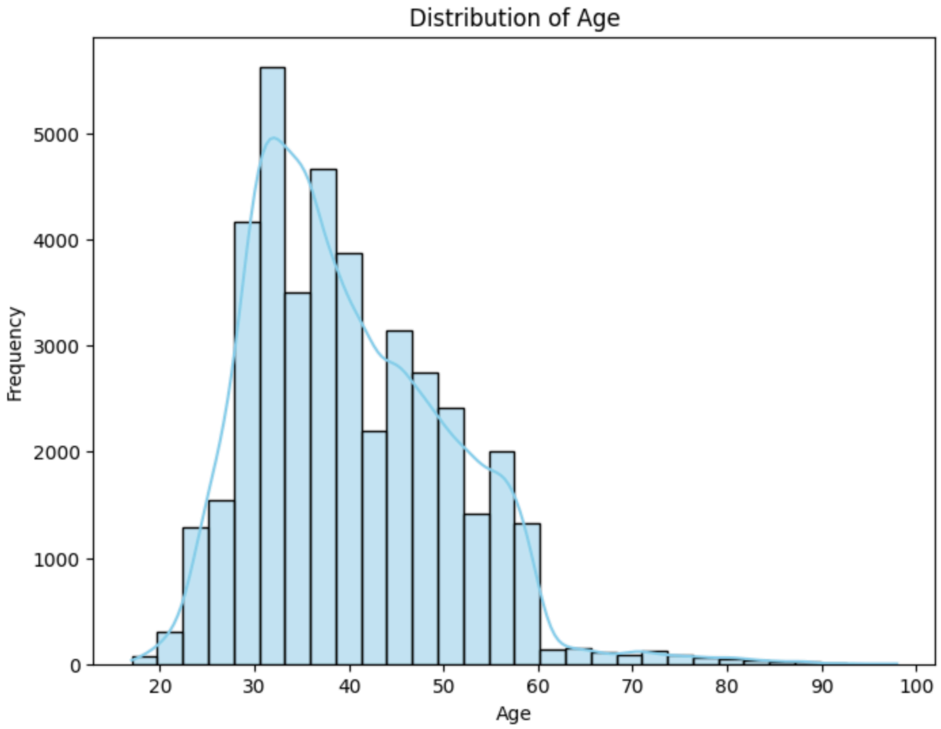
The heatmap provides an overview of the relationship between marital status and housing loan status. It reveals that married customers form the largest proportion of individuals with housing loans (12,951 customers). In contrast, divorced and single individuals are less likely to have a housing loan. This gives an insight that financial obligations such as housing loans may impact a customer's decision to invest in long-term financial products like term deposits.



Marital Status	no	unknown	yes
divorced	2092	121	2399
married	11389	588	12951
single	5097	280	6191
unknown	44	1	35

3. Age Distribution (Histogram)

The age distribution histogram shows the age demographics of customers targeted by the marketing campaign. The age distribution is right-skewed, with most customers falling between the ages of 30 and 40. A small number of older customers (60+) are also in the dataset.



Data Cleaning and Preprocessing

After gaining an initial understanding of the dataset through visualization, the next step is to clean and preprocess the data to ensure it is in the best possible shape for machine learning models.

The dataset contains categorical values where some entries are labeled as "unknown", indicating missing information. To maintain data quality, we replace these values with NaN and remove all rows containing missing data. This reduces the dataset size from 41,188 records to 38,245, ensuring only complete records are used for training.

Several features, such as marital status, job type, and education level, are categorical and need to be converted into numerical format. We use one-hot encoding for marital status and job type, creating binary indicator variables for each category. For education level, since it has a natural order, we map categories to corresponding years of schooling:
Illiterate: 0
Basic 4 years: 4
Basic 6 years: 6
Basic 9 years: 9
High school: 12
Professional course: 14
University degree: 16

Binary variables such as housing loan, personal loan, and subscription status (y) are originally stored as "yes"/"no", so we convert them into 1/ 0 for compatibility with machine learning models. Similarly, boolean values represented as True/False are also converted to 1/0 for consistency.

Finally, after encoding categorical variables, the original "job" and "marital" columns become redundant, so I remove them to prevent unnecessary duplication. After these preprocessing steps, the dataset is fully numeric and ready for training machine learning models.

Model Selection and Training

To build the predictive models, I first prepared the data for machine learning. I split the dataset into features (X) and the target variable (y), where "y" indicates whether a client subscribed to the term deposit. Using an 80:20 ratio, I divided the data into training and testing sets. To ensure all features contributed equally to the model, I standardized them using StandardScaler, which centers the data around zero and scales it to unit variance.

Before training, I noticed the dataset was highly imbalanced—most clients did not subscribe (class 0). To address this, I applied SMOTE, a technique that creates synthetic samples of the minority class (subscribers, class 1). After balancing the classes, I proceeded to train three neural network models.

First Model: Single Layer Perceptron (SLP)

I started with a Single Layer Perceptron (SLP), a simple neural network architecture with only one output layer. The model's architecture consists of a single sigmoid output unit, which is typically used for binary classification tasks. The input layer is directly connected to the output, with no hidden layers between them, making it essentially equivalent to logistic regression.

Then I compiled the model using the Adam optimizer, which adjusts the learning rate during training for more efficient convergence, and the binary cross-entropy loss function, which is suitable for binary classification. I trained the model for 100 epochs, using a batch size of 32 and setting aside 20% of the data for validation during training to track its generalization ability.

The SLP model achieved an accuracy of 83.66% on the test set. While this performance was not bad, the model's linear nature limited its ability to capture complex patterns in the data. For instance, it had trouble identifying subscribers accurately. This indicated that a better model might be necessary to improve performance and capture non-linear relationships between features.

Second Model: L2-Regularized Multilayer Perceptron (MLP)

Next, I designed a deeper network, the L2-Regularized Multilayer Perceptron (MLP), which consists of two hidden layers with 128 and 64 units. The hidden layers use the ReLU activation function, which helps the model learn non-linear relationships by allowing for the activation of neurons only when the input is positive. I also applied L2 regularization to the weights of each hidden layer using the regularizers.l2(0.01) term. This discourages large weights, reducing the risk of overfitting and promoting generalization.

Training this model for 100 epochs improved the performance to 85.10% accuracy, which was better than the SLP. The additional layers allowed the model to capture more complex relationships between the features, such as interactions between age and housing loan status, influencing subscription decisions.

Third model: MLP with Tanh Activation
For comparison, I tried another version of the Multilayer Perceptron (MLP), but this time I replaced the ReLU activation function with the tanh activation function in the hidden layers. The tanh function squashes the outputs to the range between -1 and 1, which can be useful for centering data, especially if features have negative values.

However, the MLP with Tanh activation performed poorly, achieving only 75.41% accuracy, lower than both the SLP and L2-MLP models. The training process also showed unstable learning, as evidenced by fluctuating validation loss. The tanh activation suffers from vanishing gradients when the input values become extreme. This makes it difficult for the optimizer to update weights effectively, especially in deeper layers. While the tanh function works well in some cases, ReLU proved to be more suitable for this dataset, as it allowed faster convergence and better performance.

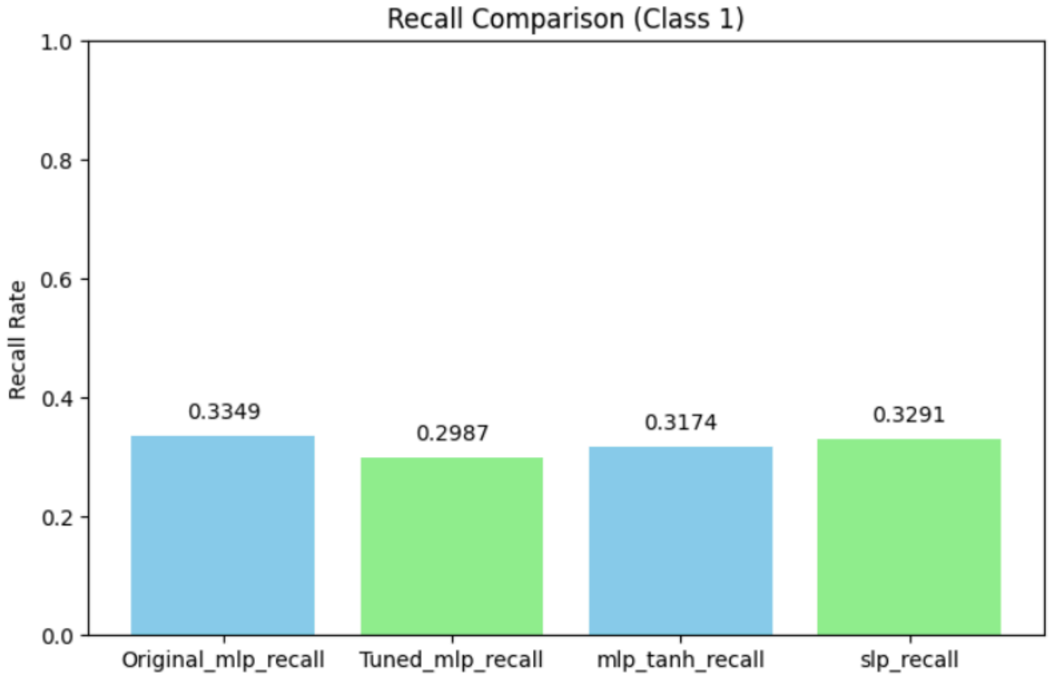
Tuning MLP model
After observing that the L2-Regularized Multilayer Perceptron (MLP) provided the best performance among the models tested so far, achieving 85.10% accuracy, I decided to fine-tune the model to further improve its performance.

The first step was scaling the features using StandardScaler to ensure all input features had the same scale, helping the model converge more efficiently. Next, I addressed class imbalance with SMOTE, generating synthetic examples for the minority class to improve the model's ability to distinguish between classes. I then modified the network architecture by adding more hidden layers and applying L2 regularization to prevent overfitting. I added Batch Normalization to stabilize training, and I also added Dropout layers to further reduce overfitting. For optimization, I used the RMSprop optimizer with a learning rate of 0.001 and incorporated ReduceLROnPlateau and EarlyStopping callbacks to adjust learning rates and prevent overfitting during training.

After training for 500 epochs, the model achieved 85.79% accuracy on the test set, confirming that the combination of these techniques significantly improved the model's performance.

Further Comparison
Given that the task involves predicting whether customers will subscribe to term deposit, recall rate becomes a critical metric. I want to ensure that the model captures as many of these potential "true positives" as possible, even at the expense of precision. For this reason, I decided to compare the recall rates of the different models (Original MLP, Tuned MLP, MLP with Tanh, and SLP) to evaluate their ability to identify customers who might subscribe.

To calculate the recall for each model, I used the recall_score function from sklearn library, setting the pos_label=1 to focus on Class 1 (subscribers). This allowed me to compare how well each model detected the positive class. I then visualized the recall rates for all four models using a bar plot to make the comparison clearer.



The results showed that the Original MLP model achieved the highest recall rate at 0.3349, followed by the Single Layer Perceptron (SLP) with a slightly lower recall rate of 0.3291. The MLP with Tanh activation and the Tuned MLP model showed even lower recall rates of 0.3174 and 0.2987, respectively. These results indicated that while the tuned model improved accuracy, it was less effective at capturing the true positives compared to the original MLP model. This was likely due to the more complex tuning process, which may have inadvertently reduced the model's ability to prioritize recall, possibly overfitting or generalizing in a way that reduced its focus on identifying potential subscribers.

Choosing Final Model

When choosing the best model, the decision should be driven by the specific objective at hand. If the goal is to maximize overall accuracy, then the model with the highest accuracy would be the best choice. However, in our case, where the business objective is to identify as many potential customers (subscribers) as possible, I must consider recall rate as a key factor. While the fine-tuned model improved accuracy, it did so at the expense of recall. Therefore, in this scenario, it is more important to select the model with a higher recall rate, ideally one that balances both recall and accuracy. Based on this, the Original MLP model, with its relatively higher recall rate, would be the most suitable choice, as it better aligns with the objective of capturing more potential subscribers, even if it means accepting a slight trade-off in overall accuracy.

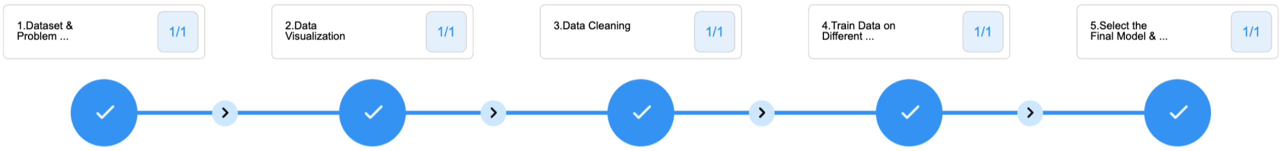
Project IPYNB file OneDrive Download Link:

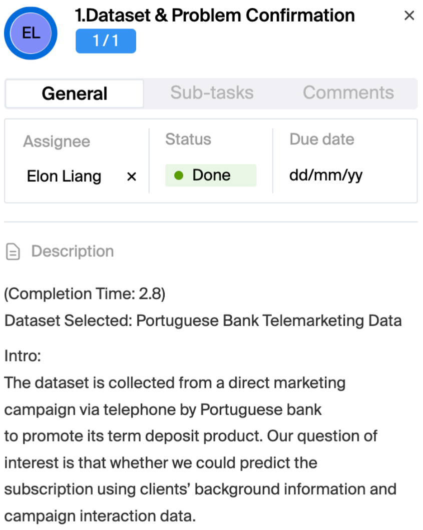
MSIN0097_Individual_Coursework

Dataset Download Link:

https://uwmadison.app.box.com/s/zbt12voi80tk3n7nwwlpfp0y5y3jz4kr

Appendix(Fractal)





1.Dataset & Problem Confirmation

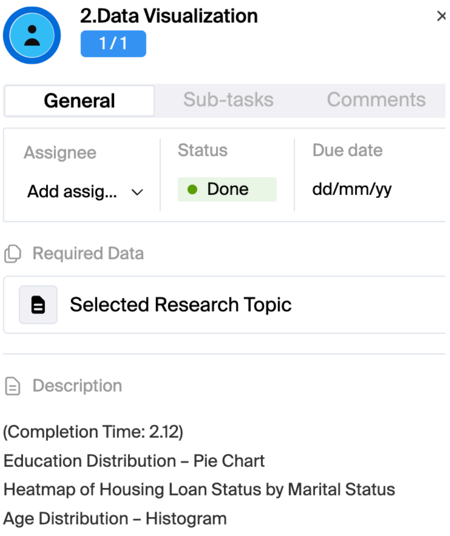
1 / 1

Assignee	Status	Due date
Elon Liang ×	● Done	dd/mm/yy

Description

(Completion Time: 2.8)
Dataset Selected: Portuguese Bank Telemarketing Data

Intro:
The dataset is collected from a direct marketing campaign via telephone by Portuguese bank to promote its term deposit product. Our question of interest is that whether we could predict the subscription using clients' background information and campaign interaction data.



2.Data Visualization

1 / 1

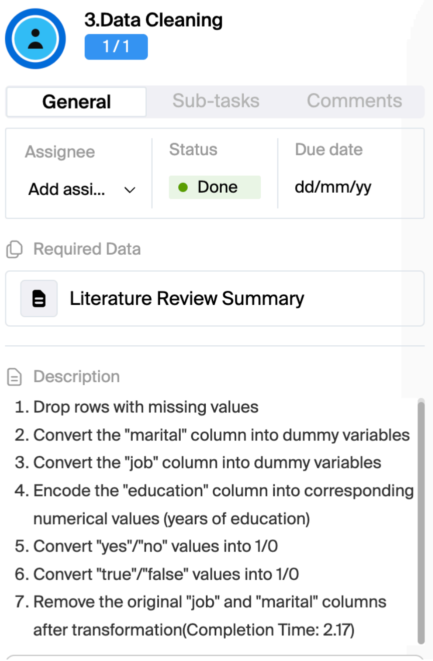
Assignee	Status	Due date
Add assig... ⌄	● Done	dd/mm/yy

Required Data

Selected Research Topic

Description

(Completion Time: 2.12)
Education Distribution – Pie Chart
Heatmap of Housing Loan Status by Marital Status
Age Distribution – Histogram



3.Data Cleaning

1 / 1

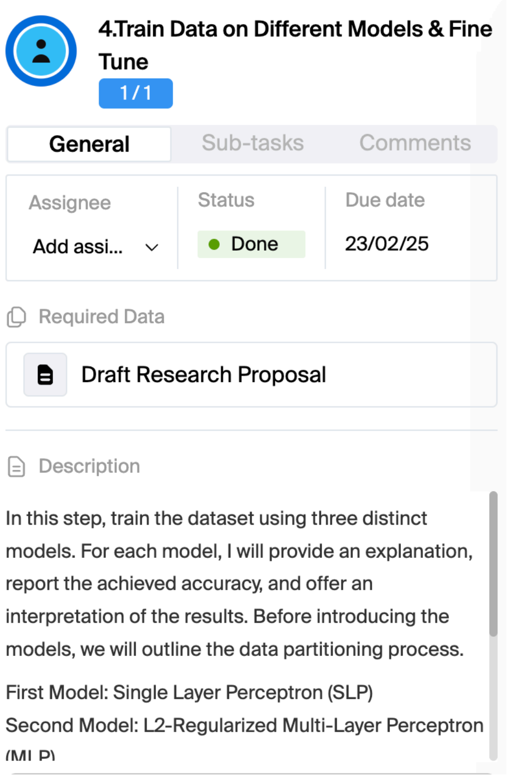
Assignee	Status	Due date
Add assi... ⌄	● Done	dd/mm/yy

Required Data

Literature Review Summary

Description

1. Drop rows with missing values
2. Convert the "marital" column into dummy variables
3. Convert the "job" column into dummy variables
4. Encode the "education" column into corresponding numerical values (years of education)
5. Convert "yes"/"no" values into 1/0
6. Convert "true"/"false" values into 1/0
7. Remove the original "job" and "marital" columns after transformation(Completion Time: 2.17)



4.Train Data on Different Models & Fine Tune

1 / 1

Assignee	Status	Due date
Add assi...	● Done	23/02/25

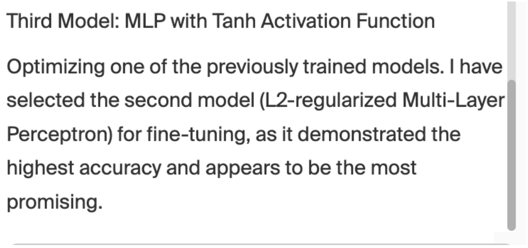
Required Data

Draft Research Proposal

Description

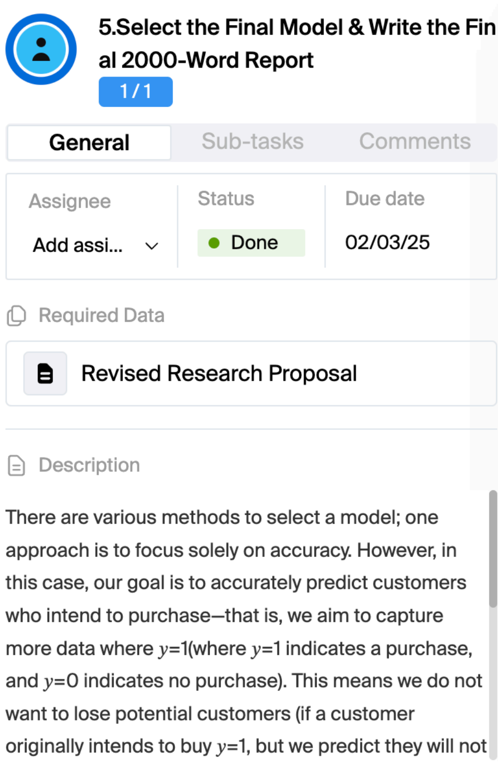
In this step, train the dataset using three distinct models. For each model, I will provide an explanation, report the achieved accuracy, and offer an interpretation of the results. Before introducing the models, we will outline the data partitioning process.

First Model: Single Layer Perceptron (SLP)
Second Model: L2-Regularized Multi-Layer Perceptron (MLP)



Third Model: MLP with Tanh Activation Function

Optimizing one of the previously trained models. I have selected the second model (L2-regularized Multi-Layer Perceptron) for fine-tuning, as it demonstrated the highest accuracy and appears to be the most promising.



5.Select the Final Model & Write the Final 2000-Word Report

1 / 1

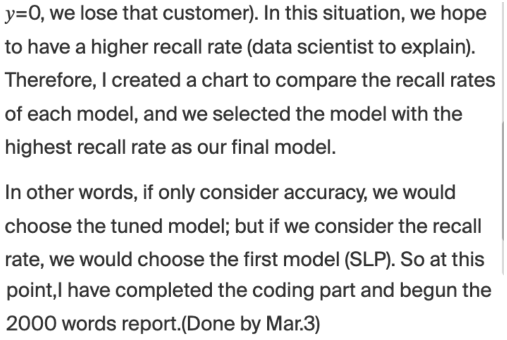
Assignee	Status	Due date
Add assi...	● Done	02/03/25

Required Data

Revised Research Proposal

Description

There are various methods to select a model; one approach is to focus solely on accuracy. However, in this case, our goal is to accurately predict customers who intend to purchase—that is, we aim to capture more data where y=1(where y=1 indicates a purchase, and y=0 indicates no purchase). This means we do not want to lose potential customers (if a customer originally intends to buy y=1, but we predict they will not



y=0, we lose that customer). In this situation, we hope to have a higher recall rate (data scientist to explain). Therefore, I created a chart to compare the recall rates of each model, and we selected the model with the highest recall rate as our final model.

In other words, if only consider accuracy, we would choose the tuned model; but if we consider the recall rate, we would choose the first model (SLP). So at this point,I have completed the coding part and begun the 2000 words report.(Done by Mar.3)