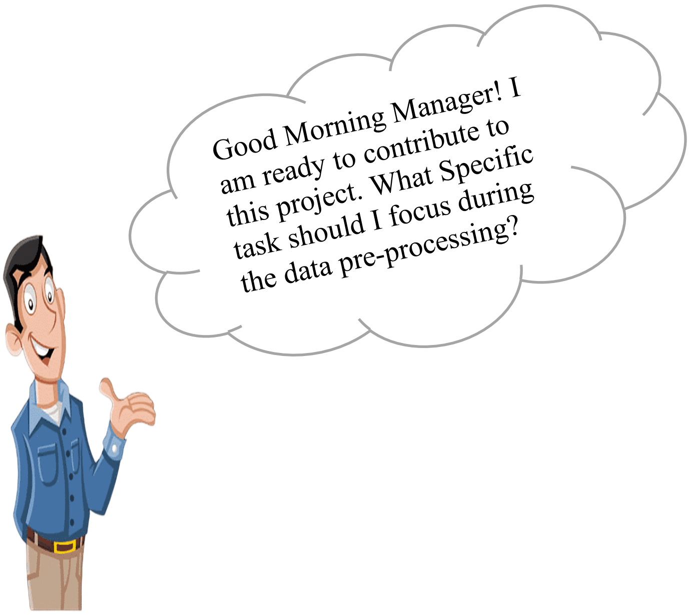
CHRIST (Deemed to be University), Bangalore – 560 029

Department of Computer Science

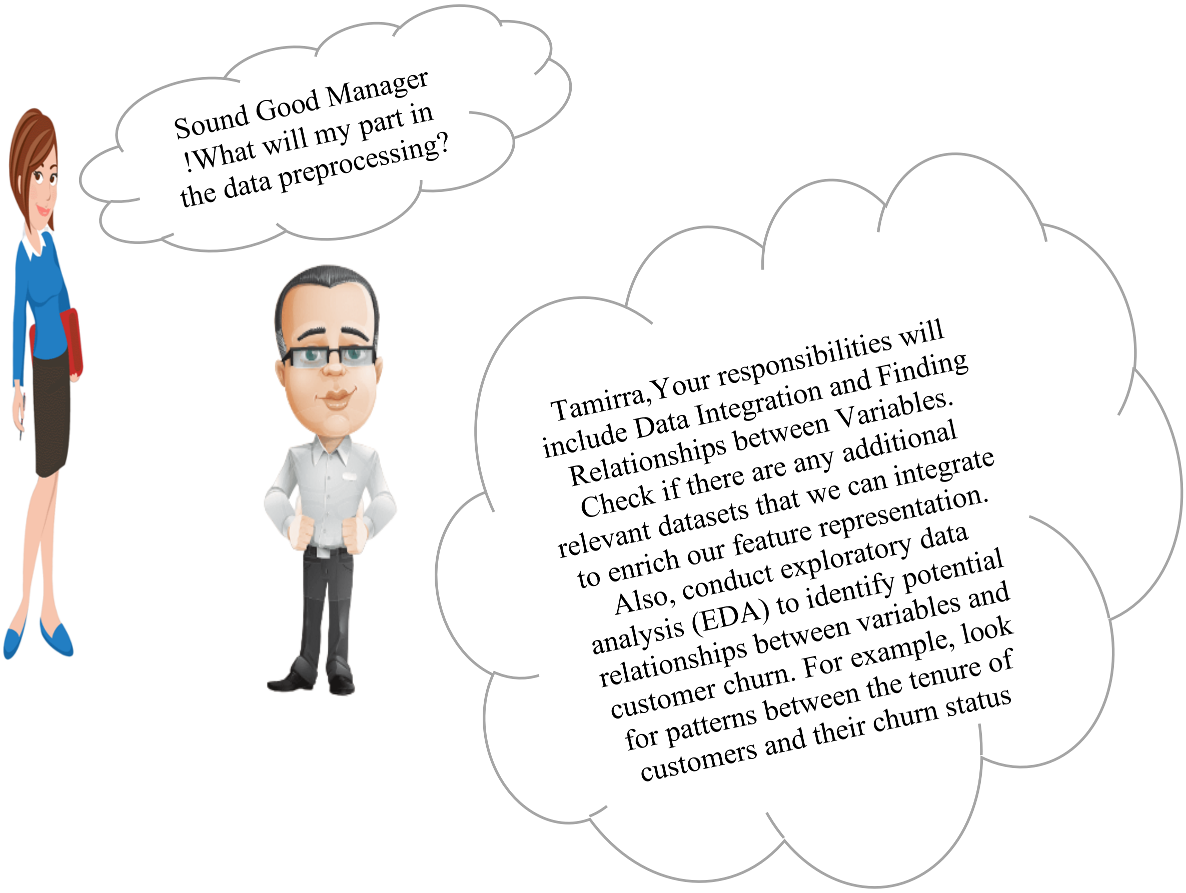
MAI171-Machine Learning, Lab 1,Date:25-7-2024

Good morning, team! As you know, we are experiencing a high churn rate in our bank, and we believe building a churn prediction model through data preprocessing can help us tackle this challenge. I have a task for all three of you: Rithul, Tanila, and Tamirra.

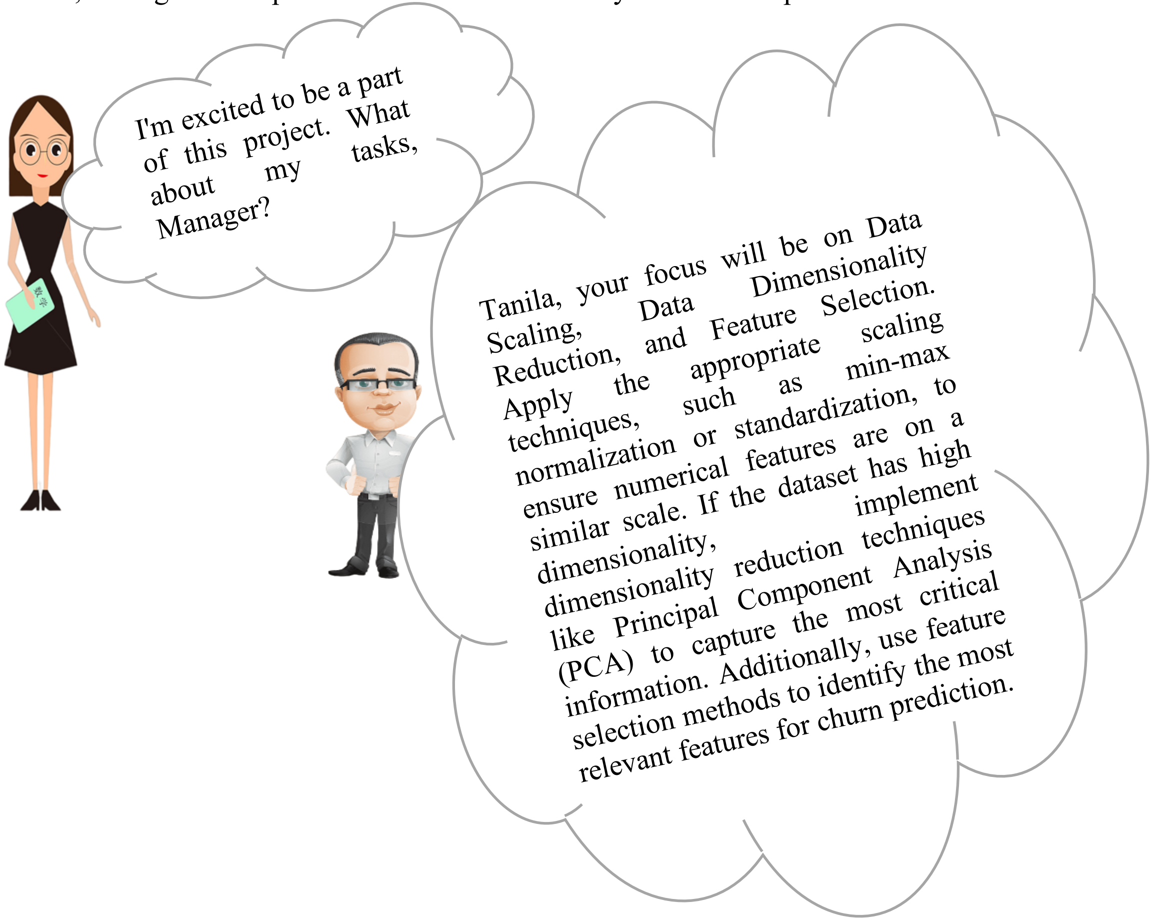

Good Morning Manager! I am ready to contribute to this project. What Specific task should I focus during the data pre-processing?

Manager: Great, Rithul! To ensure equal distribution of work, here's how we'll divide the tasks:Rithul, you will handle the Data Cleaning and Handling Noisy Data. Start by examining the dataset for any missing values, and use appropriate methods, such as mean imputation, to handle them. Also, look for any noisy data or outliers that might affect our analysis and apply binning or outlier detection techniques to handle them.

Rithul: Got it, Manager! I'll make sure to clean the data and address any noisy data.


Sound Good Manager !What will my part in the data preprocessing?
Tamirra,Your responsibilities will include Data Integration and Finding Relationships between Variables. Check if there are any additional relevant datasets that we can integrate to enrich our feature representation. Also, conduct exploratory data analysis (EDA) to identify potential relationships between variables and customer churn. For example, look for patterns between the tenure of customers and their churn status

Tamirra: Understood, Manager! I'll explore additional data and analyze relationships between variables.


I'm excited to be a part of this project. What about my tasks, Manager?
Tanila, your focus will be on Data Scaling, Data Dimensionality Reduction, and Feature Selection. Apply the appropriate scaling techniques, such as min-max normalization or standardization, to ensure numerical features are on a similar scale. If the dataset has high dimensionality, implement dimensionality reduction techniques like Principal Component Analysis (PCA) to capture the most critical information. Additionally, use feature selection methods to identify the most relevant features for churn prediction.

Tanila: Thank you, Manager! I'll work on scaling, dimensionality reduction, and feature selection.

Manager: Excellent! Additionally, all three of you should collaborate on converting relevant continuous variables to discrete using binning. For example, you can group customers' ages into age categories. Moreover, create a heatmap for visualization, which will help us understand the correlation between variables and identify any potential multicollinearity.

Rithul, Tanila, and Tamirra: Absolutely, Manager! We'll collaborate and communicate throughout the process.

Manager: Great enthusiasm, team! I appreciate your dedication. To further challenge your data preprocessing skills, I'm adding one more difficult task:

The churn dataset we have might suffer from class imbalance, where the number of customers who churned (Exited = 1) is significantly lower than those who did not churn (Exited = 0). This imbalance could lead to biased model performance and affect the churn prediction accuracy.

To address this, I want you all to explore and implement techniques to handle imbalanced data. This might involve methods like:

Resampling Techniques: Use techniques such as oversampling the minority class (churned customers) or undersampling the majority class (non-churned customers) to balance the class distribution.

Synthetic Data Generation: Consider generating synthetic samples for the minority class using techniques like Synthetic Minority Over-sampling Technique (SMOTE) or Adaptive Synthetic (ADASYN) to increase the representation of churned customers.

Manager: Perfect! Remember to provide clear explanations and interpretations of the results in your final reports. Visualizations, graphs, and tables will be essential to support your analysis and make it easier for stakeholders to understand. Take your time, and let's meet again in a week to discuss your findings and insights. Good luck to all!

Rithul, Tanila, and Tamirra :Thank you, Manager! We'll do our best and keep you updated on our progress.

Evaluation Rubrics

- **Data Cleaning and Handling Noisy Data:**2 Marks
- **Exploratory data analysis (EDA) :**2 Marks
- **Feature Scaling:**2 Marks
- **Dimensionality Reduction:**2 Marks
- **Class Imblance:**2 Marks