DS 5220: Supervised Machine Learning and Learning Theory
Summer, 2024

Dr. Fatema Nafa
In Class Work 4

Deadline: 06/ 02/ 2024. 11:59 PM

Learning Objectives:

- **Understand Linear Regression:** Students will learn how to implement a linear regression model in a statistical software or programming environment.
- **Data Preprocessing:** Handling and preprocessing data for regression analysis.
- **Interpretation of Results:** Students will learn to interpret the output of regression models, understanding concepts like the regression coefficient and intercept, and their significance in the context of the data.
- **Forecasting:** Apply the model to predict future values and understand the limitations and assumptions of the model used.

Scenario: Predicting Salary Based on Various Factors

In-Class Assignment: Employee Retention Analysis and Classification

Objective

Analyze the employee retention dataset to identify key factors influencing employee retention and develop a logistic regression model to predict employee retention. This exercise will enhance your data analysis, visualization, and machine learning skills.

Instructions

Dataset Download:

Download the employee retention dataset from Kaggle
https://www.kaggle.com/datasets/giripujar/hr-analytics

Exploratory Data Analysis (EDA):

- Perform initial data exploration to understand the dataset structure, missing values, and basic statistics.
- Identify and discuss which variables might have a direct and clear impact on employee retention (i.e., whether they leave the company or continue to work).
- Use summary statistics and visualizations to support your analysis.

Data Visualization:

- Plot bar charts to show the impact of employee salaries on retention.
- Plot bar charts to show the correlation between departments and employee retention.
- Explore other relevant visualizations (e.g., histograms, box plots) to uncover insights related to retention.

Feature Engineering:

- Based on the EDA, create new features that might be useful for the prediction model.
- Justify the choice of features selected for the model.
- Consider handling categorical variables, scaling numerical variables, and dealing with missing data appropriately.

Logistic Regression Model:

- Split the dataset into training and testing sets.
- Build a logistic regression model using the variables identified in the EDA and feature engineering steps.
- Tune hyperparameters if necessary to improve model performance.

Model Evaluation:

- Measure the accuracy of the model on the test set.
- Compute and interpret other performance metrics such as precision, recall, F1 score, and ROC-AUC.
- Discuss the significance of each metric in the context of employee retention.

Advanced Analysis:

- Compare the logistic regression model with other classification algorithms (e.g., Decision Trees, Random Forest, Support Vector Machines).
- Perform cross-validation to ensure the robustness of the models.
- Conduct feature importance analysis to determine the most influential features for employee retention.
- Perform classification analysis to categorize employees into different risk levels of leaving the company (e.g., high risk, medium risk, low risk).

Northeastern University
Khoury College of Computer Sciences

DS 5220: Supervised Machine Learning and Learning Theory

Summer, 2024

Dr. Fatema Nafa

Documentation:

- Document your entire analysis process, including the rationale for each step.
- Provide detailed explanations and justifications for your choices in data preprocessing, feature selection, and model building.
- Summarize your findings and provide actionable insights based on the analysis.

Deliverables

- A Jupyter notebook including code, visualizations, and explanations.
- A report (Overleaf) summarizing your findings, model performance, and insights gained from the analysis.

Submission

- Submit your completed work on Canvas.