GOALS AND HERISTICS OF THE EXPERIMENTAL CONFIGURATION AND TYPES

General Questions.

1. How the number if experiments were determined (24 experiments) ?

- This specific number is not determined to be 24; however, it is based on the distribution and segmentation of the experiments, which totaled 24.

- These experiments are conducted based on the original and sampled distributions.

- For the original distribution, three experiments were conducted, utilizing different test sets: University OM, University OM + Decider OM, and University OM + Decider OM + Customer Order + Online Store.

- For the sampled distributions, various percentages of positive (P) and negative (NP) instances were used: 30% P - 70% NP, 50% P - 50% NP, 60% P - 40% NP, and 80% P - 20% NP.

- Each distribution had three experiments with the aforementioned OM settings (1 OM, 2 OM, and 4 OM) for the test set, resulting in a total of 16 experiments.

- This adds up to a total of 20 experiments.

- However, the test sets used for the 16 experiments had a minimal number of samples: 31 instances for University OM, 111 instances for University OM + Decider OM, and 127 instances for University OM + Decider OM + Customer Order + Online Store.

- To address this, four additional experiments were conducted, using Traffic Controller OM as the unseen test set, which consisted of a total of 2739 instances.

- Ultimately, the number of experiments reached 24.

2. What is the Organization of 24 experiments ?

The experiments have been divided into three segments.

- **Original Segment:** This segment includes the original number of P labeled and NP labeled samples. The test sets were changed for each case. Specifically, Experiment 1 was conducted to assess the model's performance on the original distribution without any sampling operation in the dataset. The purpose was to evaluate how well the model performs on the original distribution.

- **Sampled Segment:** This segment consists of a total of 16 experiments, with four different distribution types: 30% P - 70% NP, 50% P - 50% NP, 60% P - 40% NP, and 80% P - 20% NP. The goal was to observe how P labeled samples are detected and how effective the model is on each separate distribution in identifying P labeled samples. The main motivation behind sampling was the significant difference in the number of P labeled samples compared to NP labeled samples. These distribution-based experiments helped determine the most effective distribution for identifying P labeled samples.

- **Selective Oversampled Segment:** This segment includes a total of four experiments, where a single OM (Traffic Controller) was used on four different distributions: 50% P - 50% NP, 60% P - 40% NP, and 80% P - 20% NP. The purpose was to understand the most effective distribution based on a large test set. The first distribution was not utilized in this segment because it did not perform well in identifying P labeled samples. The focus here was on selective oversampling to improve the model's performance.

3. What distinguishes each type of experiment.

- The experiments conducted in the Original segment were based solely on the original OM data and the distribution of P and NP labeled samples. These experiments provided the actual numerical representation of P and NP samples, where the NP samples were significantly larger in number compared to the P labeled samples.

- In the Sampled segment, the experiments were based on increasing the percentage of P labeled samples according to the specified distribution. For example, in the 30% P - 70% NP distribution, the original distribution of P labeled samples was increased to 30% through oversampling, while NP labeled samples represented 70% of the data. The same oversampling technique was applied to the other distributions. It's important to note that while applying oversampling, the total number of instances remained the same. Only the distributions changed to maintain consistency and stability in the training set for each distribution.

- In the Selective Oversampled segment, the experiments used the same training set as the previous 20 experiments, but the testing set was different. This introduced a different type of testing set, providing better insight into how the distributions performed on a large unseen test set. It aimed to understand the effectiveness of the distributions in a more realistic scenario.

Selective Questions.

1. Goals and heuristics behind sampling distribution.

- In the NP 70% - P 30% distribution:

  Goal: The goal was to understand the impact of P labeled samples in the dataset when they represent a significant number.

  Heuristic: Since the original number of P labeled samples was very limited (483 out of 30,971 samples), the heuristic was to increase the number of P labeled samples to a significant portion. This allowed for a better understanding of the impact of P labeled samples in the dataset.

- In the NP 50% - P 50% distribution:

  Goal: The goal was to understand the impact of P labeled samples in the dataset when they represent an equal number.

  Heuristic: The heuristic was to create an equal distribution of P labeled samples in the dataset. Although the P labeled samples were oversampled (using random oversampling), having an equal distribution helped in identifying a significant number of P labeled samples.

- In the NP 40% - P 60% distribution:

  Goal: The goal was to understand the impact of P labeled samples in the dataset when they represent a higher portion compared to NP samples.

  Heuristic: The heuristic behind this distribution was to increase the proportion of P labeled samples in the dataset. This distribution helped in distinguishing the P labeled samples from the dominant NP labeled samples in the dataset.

- In the NP 30% - P 70% distribution:

  Goal: The goal was to understand the impact of P labeled samples in the dataset when they represent a higher portion compared to NP samples and when they dominate the NP labeled samples.
  Heuristic: This distribution aimed to examine the impact of a dominant distribution of P labeled samples in the dataset. It helped in classifying and identifying P and NP labeled samples, particularly when the P labeled samples dominated the entire dataset.

- In the NP 20% - P 80% distribution:

  Goal: The goal was to understand the effectiveness of NP labeled samples in the dataset when the dataset is highly dominated by P labeled samples.

  Heuristic: In this distribution, the expectation was to see the dominance of P labeled samples. However, detecting NP labeled samples despite their lower proportion demonstrated the accuracy and effectiveness of the model in identifying both P and NP labeled samples