Math 189 Project Proposal

Authors: Sean Perry, Sahana Narayanan, Yujin Lee, Beatrice Fernandez

Focus

The use of Artificial Intelligence and Large Language Models is rising at a rapid rate. From generating code to creating summaries, LLMs such as ChatGPT can complete many of our everyday tasks and answer questions about many different topics. Our group decided to investigate the question of how ChatGPT could aid us when it comes to working on data science projects, specifically when it comes to generating synthetic data that we can use to train models.

Our study will be focused on two questions:

1. How well can ChatGPT fill in missing data?
2. Can the output of ChatGPT be used to train downstream classification models?

For the first part of the experiment, we want to learn if ChatGPT can reliably predict missing data values. For this, we will consider the datasets mentioned below, one with multiple missing values and the second without missing values where we will manually remove values to compare ChatGPT's predictions with the actual values. From this, we will consider, what kinds of missingness ChatGPT tends to be better at imputing? What kinds of text prompts improve ChatGPT's ability to impute? For our second main question, we aim to experiment with ChatGPT for complete missing data in a fresh dataset, comparing its performance against traditional techniques or leaving the data incomplete. This evaluation seeks to gauge the quality of training data generated by ChatGPT.

Data Background

Our project will center on three datasets acquired from Kaggle and sourced from reputable platforms: the HR Analytics Job Change of Data Scientists dataset from Analytics Vidhya, the International Students Demographics dataset from Open Doors Data, and the New York City Airbnb Open Data sourced from Inside Airbnb. The HR Analytics dataset contains missing values for attributes such as Gender, Enrolled University, Education Level, and Discipline; the International Students dataset exhibits missing values for US Students, Undergraduate, Graduate Degree, and Non-Degree categories. Meanwhile, the Airbnb dataset serves as a complete reference. Through analyzing these datasets, we aim to evaluate ChatGPT's efficacy in filling missing data and enhancing the overall completeness and accuracy of diverse datasets.

Analysis

For the first experiment we will conduct a test-train split, select a column in the dataset, simulate missing data in that column, and create a text prompt designed to help fill in the missing data

based on the information from the rest of the row, use the template filling in known metadata for each row's missingness using ChatGPT 3.5 API, record evaluation metrics (accuracy, precision, and recall), and if time permits use tradition techniques for estimating missing data. Upon the completion of the experiment, we will have a set of predicted values and true values in the test dataset. We can then use hypothesis testing to compare the sample of the labels and the sample of the predicted values and say something about ChatGPT's performance at filling in data. If we have time for testing traditional techniques for imputation, we can also do a test to see if the distribution of correct samples is better for ChatGPT than for traditional techniques.

For the second experiment we will consider the best-performing template style (from the first experiment), use the dataset with missing data, split the dataset based on a random prop of rows without missing data (for fairness), use ChatGPT template, traditional techniques or nothing to the resulting data, train an upstream model (such as KNNs, SVMs) on possible classification tasks for the dataset based on the newly filled dataset and control dataset, and report the classification results for each dataset and upstream model. Thus this will look at what distributions of predictions for the upstream task, and see if the model had more correct predictions with ChatGPT imputation. So we would want to see that the true performance of a model trained on ChatGPT augmented dataset distribution is better than the true performance of a dataset just filtered for nonmissing rows or imputation with traditional techniques.

Expected Outcomes

The anticipated outcomes of the study encompass a comprehensive understanding of ChatGPT's effectiveness in addressing missing data and its potential for enhancing downstream classification model training. By systematically assessing ChatGPT's performance across diverse datasets, the study expects to discern patterns in its handling of missingness and identify optimal text prompts for improved imputation accuracy. Moreover, through comparison with traditional techniques, the study aims to elucidate how ChatGPT surpasses existing methods in filling in missing data. Additionally, the investigation is expected to reveal the impact of ChatGPT-generated training data on the performance of upstream classification models, clarifying whether such data can produce better classification results compared to traditional methods or incomplete datasets. Ultimately, this research strives to provide empirical insights that inform the integration of ChatGPT into data preprocessing pipelines and improve the efficiency and accuracy of machine learning workflows.