Research Experience Notation (REXP) – Summer 2023 ABSTRACT

Name:	**Sid Lamsal**
Major:	**Computer Science and Data Analytics**
Anticipated Grad Year:	**2025**
Email:	**lamsals@dickinson.edu**
Phone:	**717-317-8710**

Abstract:

Please provide an Abstract (250-word limit) that objectively reports the outcome of your completed research experience. You should follow the conventions of scholarship in your field of research.

Generally, an Abstract should briefly describe the research question and the methods used to investigate the question, and should summarize the results of the investigation as well as the significance of those results.

Do not include remarks about your subjective personal experience of the work. Save those for the Reflective Essay.

Our research aimed to create an interpretable preprocessing method for classification using covariate matching to improve neural network performance. We utilized two customer churn data sets from a telecom company in California. Our goal was to use variables related to customer churn (covariates) to pair customers with similar telecom services but opposite churn behavior, i.e., “hard-to-predict” pairs.

We found it optimal to use logistic regression probabilities to label "hard-to-predict" pairs. If customer A churned with a low probability but customer B did not, customer A is “hard-to-predict” and paired with customer B. We then prioritized hard-to-predict pairs by adjusting the neural network’s loss function, changing the sampling order, and using specialized neural networks. The optimal method was to utilize two specialized neural networks, one of which was trained exclusively on hard-to-predict customers. The combined neural networks modeled churn data significantly better than a single neural network. We created a consolidated function which identified hard-to-predict customers, trained 2 specialized neural networks, identified hard-to-predict test set customers, and retrieved classification predictions from the appropriate neural network.

Our findings suggest that businesses can better model customer churn by identifying hard-to-predict customers and using specialized models. The models could provide insights into why hard-to-predict customers behave differently, which can be crucial for customer retention campaigns. However, the cumulative error in predicting hard-to-predict customers and subsequently predicting churn resulted in insignificant improvement in churn predictions. Moreover, the runtime of our function suggests that neural networks may not be energy efficient for simple but large-scale applications.