INDIANA DATA MINE CORPORATE PROJECT INTERIM REPORT

Kashyab Ambarani Derek Brown Ha Le Kale Menchhofer Tessa Thatcher

Department of Mathematical Sciences, Purdue University Fort Wayne

CENTRAL INSURANCE

EXCELLENCE IS OUR POLICY.™

About Central Insurance

Central Insurance Companies are a property and casualty insurance group headquartered in Van Wert, Ohio. Founded in 1876 as Central Mutual Insurance Company, the company has expanded to selling insurance products in 22 states and operates regional offices in Texas, Massachusetts, and Georgia. Additionally, Central generates $770 million in annual revenue and has 650 employees.

Fig. 1: Central Insurance's headquarters in Van Wert, Ohio.

Introduction

The topic of the corporate sponsored project is pricing insurance policies, and the goal for the team at Purdue Fort Wayne is to devise their own insurance pricing algorithms. More specifically, the team has been tasked with producing more accurate pricing algorithms for renter's insurance and insurance for condominium owners than those currently being used by a multi-million-dollar company. By leveraging data and utilizing cutting-edge statistical techniques, the team at Purdue Fort Wayne is optimistic about the results they can attain.

Overview of the Data

The data used for the project thus far has primarily been data provided by Central Insurance. The data Central Insurance provided the team at Purdue Fort Wayne concerns historical claims, policies, premiums, and billing data. The team was originally provided with seven datasets from Central—the smallest of which containing 3,862 rows of 12 variables and the largest containing 6,292,205 rows of 4 variables. Data obtained from third-party sources has also been considered for the project, including census and geographical data.

Computational Resources

Most of the tasks that have been carried out for the project thus far have been conducted using applications associated with the Brown Community Cluster. The Brown Community Cluster is a system of hardware located at Purdue University in West Lafayette, Indiana. Operations for the project are performed on Brown Cluster servers to ensure the safe storage of sensitive corporate data.

Summary of the Progress Made

The first several weeks working on the project were spent exploring the data that was provided by Central and becoming acclimated to Purdue's Community Cluster software. The exploration of the data entailed calculating summary statistics for individual variables, generating histograms and boxplots to investigate the distributions of variables and identify outliers, and performing statistical tests for variables contained in the same dataset. The subsequent weeks were spent merging the datasets provided by Central—an undertaking that required a collaborative effort from all members of the team at Purdue Fort Wayne. The merged dataset, which contains data from all seven of the original datasets provided by Central, in its entirety is comprised of 179,158 rows of 79 variables. Once the seven datasets had been merged into a single dataset, the team at Purdue Fort Wayne further investigated the originally provided data. Having the data in its form in the single dataset enabled analyses to be more easily performed on variables that were originally contained in separate datasets. During this time, additional plots were created to better understand relationships between variables and statistical tests were carried out on sets of variables to gain insight into which variables are most influential for predicting claims. The plots that were created included scatterplots, which were produced to study the relations between variable pairs, and boxplots, which were constructed for categorical variables to study the potential influences of different levels within the individual categorical variables. The next step for the project was creating linear models pertaining to predicting claims. The work that was done during this stage produced significant knowledge on the importance of particular variables to the team. At the same time the linear models were being created, cluster methods were implemented to enable geographic factors to be appended to variables already being considered. Unforeseen technical issues inhibited the results yielded from clustering methods; although, due to the advantages gained from these methods, the team intends to continue to consider clustering methods as the project continues. The team at Purdue Fort Wayne has made much progress on this project, though much still remains to be accomplished.

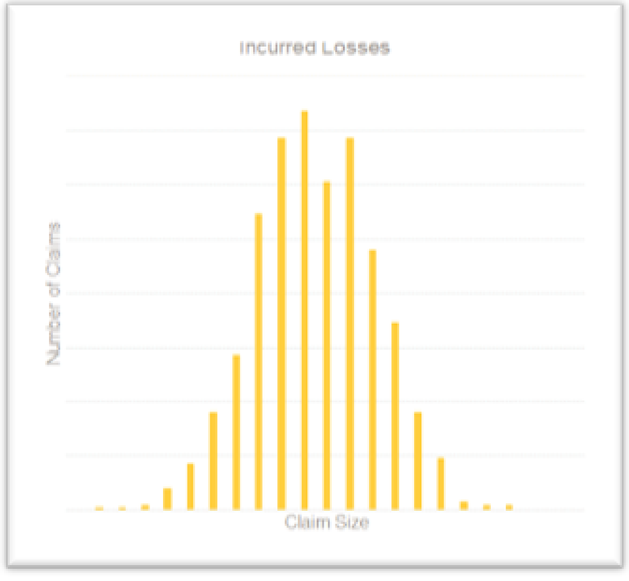

Fig. 2: A histogram depicting the severity of incurred loss for policies in which a claim was made.

Future of the Project

There is much work ahead for this project that will be carried out over the duration of the Spring 2022 semester. One of the next steps that will be taken will be working on pricing. Additionally, the pricing algorithm will need to be tested against Central's current algorithm, with the expectation that it should more accurately predict claims. The pricing algorithm will need to go through several iterations, with each revision made improving the accuracy of the algorithm. To conclude the project, the team at Purdue Fort Wayne will present their completed algorithm to Central Insurance, explaining its components and quantifying its improvements over their current algorithm.

Acknowledgements

All members of the team at Purdue Fort Wayne maintain that the experience of working on this project has been an invaluable experience. The team at Purdue Fort Wayne expresses their sincerest gratitude to Central Insurance for allowing them the opportunity to work with the professionals at their company on this project.Furthermore, the team is profoundly appreciative of Professor Mark Ward and the assistance he has provided through an introductory workshop to the Purdue Brown Server and aiding the team through difficulties encountered.

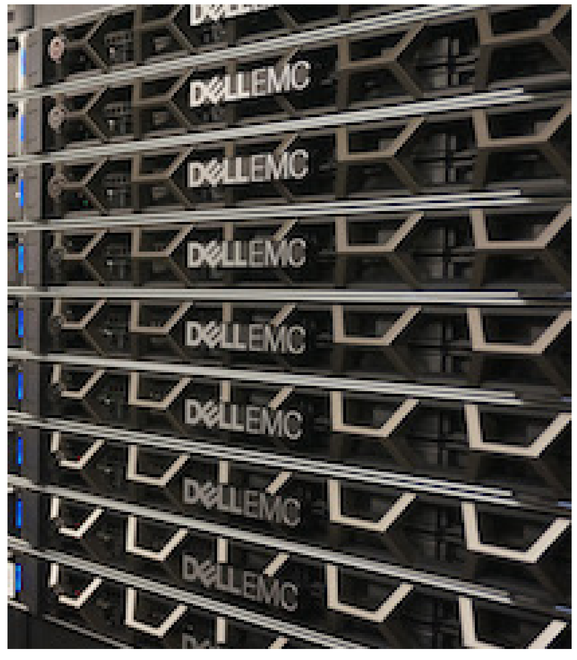

Fig. 3: A few of the many Dell compute nodes contained in the Brown Community Cluster at Purdue University.