Prudential ADS Technical Audit

Authors: Amando Xu & Rohan Mukerji
New York University, New York City, NY
Emails: acx202@nyu.edu, rm5176@nyu.edu

1 Background

The insurance industry has witnessed a remarkable transformation by integrating advanced data analytics and machine learning techniques, presenting many opportunities to revolutionize traditional risk assessment methodologies. Prudential, a prominent life insurance provider with 140 years of experience in the United States, seeks to capitalize on these advancements by implementing an Automated Decision System (ADS) to predict risk scores for life insurance policies. Hence they started a Kaggle competition to find the best implementation of an ADS for their life insurance quote system. The solution for this technical audit was an implementation that harnesses ridge regression. The primary objective of this ADS is to streamline the life insurance application process, which has become antiquated in a world where instant gratification and one-click shopping experiences are the norms. Currently, obtaining life insurance involves clients providing extensive information, undergoing medical exams, and waiting an average of 30 days before receiving a quote, which has deterred many potential customers.

In response to this challenge, the solution aims to help Prudential to provide a more efficient and less labor-intensive method for clients to acquire life insurance quotes while maintaining strict data privacy standards. When training the model, the goal is to accurately predict an individual's general risk without incorporating historical biases that may be present within sub-populations, thereby ensuring a fair and unbiased assessment. As the ADS seeks to optimize multiple goals, trade-offs may arise when identifying biases across various sub-populations, necessitating the selection of variables that minimize inherent bias while maximizing predictive accuracy.

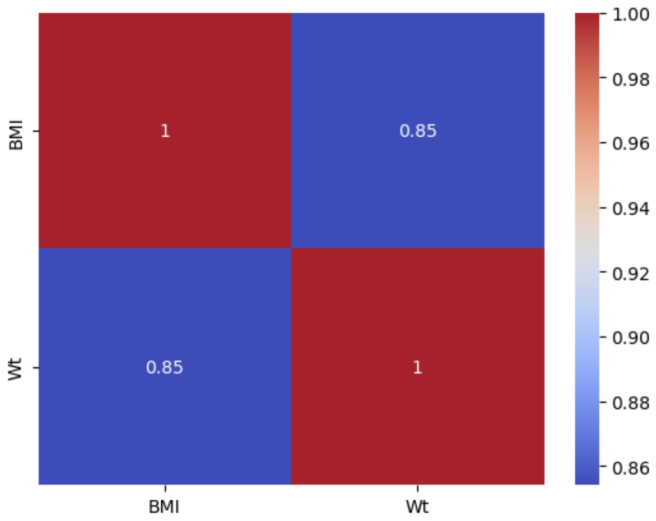
This report presents a comprehensive technical audit of the Prudential ADS, examining its effectiveness, potential implications, and areas for improvement in a rapidly evolving insurance landscape. By scrutinizing the model's performance and exploring possible trade-offs, we aim to offer valuable insights into the predictive power of the data points utilized in the existing assessment process. Ultimately, the findings from this audit will contribute to Prudential's ongoing efforts to significantly streamline the life insurance application process, fostering a more positive public perception of the industry and increasing accessibility to life insurance coverage for individuals and families across the United States.

2 Input and Output

The company meticulously collected the input data utilized by the DS and subsequently made it available through a Kaggle competition. Although the selection process for the dataset remains unspecified, it consists of 59,381 individual records, each containing a comprehensive range of 128 columns. These columns encompass various demographic, employment, insurance history, family history, and medical history information, providing a rich data source for the ADS to process and analyze. The 128 columns are further categorized into three distinct data types: 60 categorical, 13 continuous columns, and five discrete columns, while the rest are dummy variables.

The training dataset features 7 Product_Info columns, with a majority characterized as integers, one identified as a categorical variable, and one designated as a continuous data type. Notably, the age, height, weight, and BMI variables have been normalized within the dataset to facilitate analysis.

A thorough examination of the sub-populations present in the dataset reveals a particularly significant pairwise correlation between the BMI and weight variables. A heatmap has been included in the appendix to provide a comprehensive visual representation of this correlation. This heatmap also illustrates any additional correlations that may exist between the various features present in the dataset.



By rigorously profiling and understanding the input data, we can better assess the performance of the ADS and evaluate the efficacy of its ridge regression model in accurately predicting life insurance risk scores.

The output of the ADS consists of a risk score labeled as “Response,” along with the corresponding client ID. The primary purpose of this risk score is to enable Prudential’s automated system to generate precise life insurance quotes for potential and existing clients. Although the relationship between the risk score and the insurance quote remains unclear due

to the absence of explicit guidelines from Prudential, the score is a critical element in streamlining the company's processes and enhancing overall efficiency.

It is worth noting that multiple factors, including demographics, employment information, insured information, insurance history, family history, and medical history, could influence the "Response" risk score. These factors collectively contribute to generating a comprehensive risk profile for each client. However, without specific details on interpreting the risk score, it is challenging to determine if a higher score corresponds to a higher insurance quote or vice versa.

Nevertheless, adopting the ADS and its risk score output signifies a significant shift in the life insurance industry. By leveraging the power of machine learning and advanced algorithms, the system has the potential to provide more accurate, personalized, and expedited life insurance quotes, ultimately improving customer satisfaction and increasing the overall accessibility and adoption of life insurance coverage.

3 Implementation and Validation

The ADS began with preprocessing the data. They initiated the process by plotting all the categorical data in a subplot and then getting dummy variables. This is to quantify this data to be used for model training and testing. This is followed by plotting a subplot of histograms of the continuous data. These plots were performed by creating their own functions to apply to the respective data. These columns were normalized, and missing values were set to 0. These continuous data distributions were observed to be skewed; thus, the author applied Box-Cox transformation to the non-missing values while replacing the missing values with the average of the values. The discrete data was processed in the same manner as the continuous data. The Medical_Keywords 1 through 48 were made to quantify these categorical variables.

After all the preprocessing steps, the prediction model is constructed using Ridge Regression models. Before any model was predicted, several experiments were made to judge what columns should be used and determine the value of the hyper-parameter alpha by doing a train_test_split with 40000 rows for training and 10000 rows for testing. To minimize the MSE, they found that all columns would be used in the prediction model with a set alpha of 10. This is to test the mean square error by the alpha to see what the different types of data change as alpha increases.

For the prediction model, all the columns of data were concatenated and separated into train and test sets. The Ridge model was initialized with alpha equals ten and then fit into the training data. They followed by predicting values to the X_test values. In this running, the Ridge Regression Outputted a score of 0.55443. Finally, the implementation plotted the current scores in a histogram.

4 Outcomes

In this section, we analyzed the performance and fairness of the ADS model across different subpopulations based on Age, Height, Weight, and BMI. The performance was evaluated using Root Mean Square Error (RMSE). The following metric analysis is implemented in the attached .ipynp file, which shows the steps toward the RMSE metric subpopulation analysis, class imbalances, and fairness metrics.

The RMSE values for the normalized age subpopulation values were as follows: 1.6758 for Young Adults, 1.8275 for Middle-aged individuals, and 2.2261 for Seniors. Notably, the class distribution within each age group revealed significant imbalances in the 'Response' variable. In both the Young Adults and Middle-aged subpopulations, class 8 had the highest proportion. Conversely, class 6 had the highest proportion within the Seniors subpopulation. These findings indicate that the model performs most effectively for young adults and demonstrates relatively poorer performance for seniors.

The RMSE values for the normalized height subpopulation values were as follows: 1.8752 for Short individuals, 1.8997 for Medium individuals, and 1.9214 for Tall individuals. These findings suggest that the model performs similarly across height, with slightly better performance observed for short individuals. Additionally, the class distribution analysis revealed imbalances in the 'Response' variable within each height group, with class 8 having the highest proportion in all three categories.

The RMSE values for the normalized weight subpopulation values were as follows: 1.8613 for Underweight individuals, 1.9312 for Normal Weight individuals, and 1.8693 for Overweight individuals. These results indicate that the model performs best for underweight individuals and shows relatively poorer performance for normal-weight individuals. Furthermore, the class distribution analysis exhibited imbalances in the 'Response' variable within each weight group, with class 8 having the highest proportion in both the underweight and normal-weight categories. In contrast, class 6 had the highest proportion in the overweight category.

The RMSE for the normalized BMI subpopulation values were as follows: 1.8737 for Underweight individuals, 1.9269 for Normal individuals, 1.9723 for Overweight individuals, and 1.8160 for Obese individuals. These findings suggest that the model performs best for the obese category and demonstrates relatively poorer performance for the overweight category. Moreover, the class distribution analysis demonstrated significant imbalances in the 'Response' variable within each BMI group, with class 8 having the highest proportion in the underweight, normal, and overweight categories. In contrast, class 6 had the highest proportion in the obese category.

Our performance analysis revealed that the ADS model's performance varies across different subpopulations based on Age, Height, Weight, and BMI. The class distribution showed

imbalances within each group, which could affect the model's performance and should be considered when interpreting the results. All RMSEs for the subpopulations created scored consistently over 1. Further research and improvements are needed to address these biases and enhance the model's performance for all subpopulations.

Regarding this next step, our goal is also to analyze the fairness of the ADS using ridge regression concerning different sub-population groups. We have chosen demographic parity and equalized odds as our fairness metrics. Demographic parity ensures that the positive prediction rate is similar across all subpopulations, while equalized odds ensure that both true positive rates and false positive rates are equal across subpopulations. These metrics were selected because they help us assess whether the ADS treats different demographic groups similarly, thus promoting fairness.

Our demographic parity analysis reveals disparities in the positive prediction rate across the subpopulations. For age, seniors have a higher positive prediction rate (0.781) compared to young adults (0.579) and middle-aged individuals (0.079). In height, short individuals have a higher positive prediction rate (0.364) than medium (0.092) and tall individuals (0.256). For weight, underweight (1.022) and overweight individuals (1.082) have higher positive prediction rates than normal-weight individuals (0.016). Lastly, in terms of BMI, underweight (1.075) and obese individuals (1.219) have higher positive prediction rates than normal (0.355) and overweight individuals (0.213). These disparities suggest that the ADS may not be fair with respect to demographic parity.

As for additional advanced methods, we propose employing the SHAP method to analyze the performance of the ADS. SHAP is a powerful and flexible method for interpreting machine learning model predictions. It leverages game theory to provide local explanations for individual instances, enabling a deeper understanding of how each feature contributes to specific predictions. The choice of SHAP is justified by its ability to assess the model’s stability, robustness, and performance across various examples.

Our suggested methodology would involve training the ridge regression model on a split of the original training dataset and initializing the SHAP explainer with the trained and the training dataset split. We would then select an instance from the test set to interpret and compute the SHAP values for that instance. Finally, we would visualize the SHAP values using a waterfall plot, which would clearly depict the feature contributions to the single prediction and make it easier to interpret the model’s behavior.

By applying this methodology, we expect to gain insights into the model’s decision-making process and identify potential biases or shortcomings. This additional analysis should prove helpful in improving the model’s performance, addressing fairness issues, and ensuring the ADS's stability and robustness.

5 Summary

Overall, the data consistently scored relatively high root mean square errors that were over 1 across the subpopulations. While there were cases of class imbalance, we can say the data for predictive analysis is entirely appropriate for that outcome. Through our assessment and analysis of the implementation, we believe the implementation is accurate and fair based. It is not as robust as it could be, as it implements a single Ridge Regression model. As there was consistently low RMSE across the subpopulations, the Prudential corporation would find this favorable as strong predictive power exists. With the results of the demographic parity, stakeholders that prioritize bias and fairness with data would also be satisfied. This ADS included only a single type of regression for its implementation. While it was robust and accurate, we believe an ADS worth deploying would be more robust in assessing the data on a larger scale with multiple different models. For this ADS, we found the pre-processing methods account for the necessary means for modeling. However, the significant improvement would be in the complexity of the model selection.