On The Effect of Polygenic Risk Scores and other Non-Genetic Factors on Life Insurance Underwriting Decisions

Emiliano Careaga[*] Jameson Augustin[†] 2025

Abstract

This study investigates the effect of polygenic risk scores and other variables on underwriting decisions in life insurance using machine learning. Using simulated data of 10,000 individuals, we developed logistic regression and random forest models to analyze the impact of each variable. Traditional variables and their limitations were described in order to contextualize the original methods for life insurance underwriting. Other genetic methods and their limitations excluding polygenic risk scores were outlined with the intention of comparing an emerging concept with various traditional genetic analytical methods. Data was generated and processed using machine learning techniques to ensure reasonable results. However, our multicollinearity analysis revealed important limitations: the synthetic PRS showed multicollinearity with other variables (VIF > 10) and correlations with family history variables that were artifacts of the data generation process. These analyses demonstrated that polygenic risk scores have a significant effect on life insurance underwriting decisions by enhancing the accuracy of the risk-prediction models. Feature importance charts showed that the more accurate model (the random forest, as confirmed by comparison tables) gave greater weight to variables related to polygenic risk scores and their interactions. However, the model also considered several traditional variables with substantial weight, including premium cost, underlying condition, age, and health risk score. This research offers a valuable perspective into the relationship between polygenic risk score and individual characteristics, highlighting the value of integrating genetic variables with traditional variables, while acknowledging the challenges of implementing synthetic PRS. These findings provide evidence supporting the potential application of polygenic risk scores in risk-prediction models, though real-world implementation would require addressing the methodological limitations identified. By increasing the accuracy of risk-prediction models, current conflicts such as adverse selection and information asymmetry could potentially be resolved with the addition of certain policies and proper implementation of true genetic risk scores.

[*] High School Student at Maine South High School

[†] Ph.D. Graduate from The University of Georgia

I. Introduction

When calculating the amount of premium that an insured will need to pay, insurance companies can have a challenging time accurately allocating an equitable amount. This is due to the variety of factors that can contribute to an estimate of an individual's risk. In order to address this problem, several more complex genetic factors, beyond basic demographic factors, have been incorporated into life insurance underwriting decisions. Nonetheless, it remains unclear how reliable or important these new factors truly are. We hypothesize that integrating these new factors with traditional ones will significantly improve the accuracy of risk assessment.

Formerly, life insurance companies solely considered demographic factors when making under writing decisions. Although these factors are indispensable, in isolation, they overlook several other critical factors that can drastically change the probability of insureds receiving coverage. Moreover, the implementation of these critical factors, including health and lifestyle, is a crucial milestone in life insurance underwriting decisions. By incorporating these factors, they are able to take a further stride towards equitable premium pricing.

Advancements in technology in the 21st century allows for the consideration of further complex factors, the most significant of which includes genomics. In underwriting, genomics involves insurers using individuals' genetic data to estimate disease risk. It further enhances the ability for insurance companies to assess risk with complex analysis.

However, there are several different methods to gather genetic information. Considering that genetic testing creates a variety of complex results, insurance companies must have the capability to organize the information presented from the genetic tests into an underwriting decision variable. This process is achievable by either analyzing mutations, monogenic variants, or epigenetic changes.

In addition to analyzing single-gene mutations, polygenic risk scores provide a numerical estimate of an individual's genetic predisposition of a disease. Contrary to other methods, polygenic risk scores are calculated by accumulating the effects of many genetic variants, each of which contributes a small amount to overall risk. It is important to note that true PRS represents relative risk compared to population distributions, not absolute risk, and requires integration of thousands to millions of genetic variants from genome-wide association studies (GWAS). This study uses a synthetic approximation that may not fully capture this complexity. By considering genetic variants of varying effect sizes, polygenic risk scores help insurance companies set fairer premiums.

Nevertheless, there are negative consequences associated with implementing innovative technology. The implementation of genetic testing and other health or lifestyle technology allows for access to individuals' private information that may prefer to be kept confidential. In addition, genetic discrimination could potentially occur as insurers can deny insurance to higher-risk individuals, known as adverse selection. Whether or not insurance companies are able to access this information, information asymmetry, or when one party has more information than the other, is bound to occur. Insurance claim denials have arisen in the past due to insurance companies having a high volume of information, leaving high-risk individuals to carry the entire burden of the cost of mortality (Low et al. 1632–1635). As long as both parties have the same amount of information, polygenic risk scores provide a more accurate representation of risk, and in that manner, they help reduce adverse selection by enabling fairer and more accurate pricing. While it is not perfect, it helps reduce the risk of financial loss for both insurers and policyholders.

In this study, we aim to analyze several individualistic factors and their polygenic risk score to determine their effect on the insured's premium price, and the insurer's decision on whether to provide insurance by utilizing life insurance underwriting variables with realistic distributions to demonstrate the magnitude of impact each variable had. Our study will generate visual representations to present insurance companies decisions, substantiated by variable importance. The factors that will be considered include polygenic risk scores, age, gender, race, underlying conditions, region, income, father's health history, mother's health history, lifestyle, occupation, insurance type, smoking status, drinking status, and dietary choices. By analyzing each model generated, we seek to understand the correlation between key underwriting variables and insurers' coverage decisions, while evaluating the added predictive value of polygenic risk scores.

The paper is organized as follows: Section 2 describes the literature review. Section 3 presents the data. Section 4 outlines the methodologies. Section 5 discusses the implication of the results. Section 6 highlights the conclusions that can be drawn from the results.

II. Literature Review

2.1 Traditional Variables of Life Insurance

Traditional variables of life insurance often include demographic factors, such as age, sex, and occupation. The absence of technology limited the amount of information that an insurance company could access. Consequently, standardized premium rates were used to minimize potential cost variability. In addition to demographics, health, lifestyle, behavioral, and socioeconomic factors were incorporated into underwriting decisions, enabling insurance companies to create differentiated premium price levels. While these variables have augmented success, they fail to incorporate several other complex factors associated with risk.

2.2 Limitations of Variables

2.2.1 Limited Scope

Traditional underwriting methods focused on relatively few variables due to inadequate technology. Significant factors in assessing an individual's risk, such as susceptibility of developing a disease, are disregarded. As a result, there is less statistical data available to be assessed, therefore increasing the probability of misrepresenting an individual's authentic level of risk.

2.2.2 Static Nature

As measuring changes in individuals' characteristics was particularly demanding before technological evolution, traditional variables are susceptible to functioning as exclusively fixed factors. By relying only on fixed factors, premium pricing remains constant while an insured's level of risk fluctuates. Consequently, either insurers or insureds bear a greater burden than the other.

2.2.3 Bias Potential

Because traditional variables are largely demographic and static, they may introduce bias or lead to discriminatory outcomes. For example, gender, race, and income are all generalized variables that can classify an individual in a pool with higher or lower risk. By only incorporating these determinants, insurance companies are compelled to infer assumptions regarding an individual, further misrepresenting an individual's level of risk.

2.3 Different Methodologies Concerning Genetic Testing

As technological advances were made, complex variables, including parent health status and genomic analysis, were implemented to further enhance accurate premium pricing. Due to an influx of information, insurers were capable of providing flexible premiums.

Although genomic analysis can be interpreted in several ways, it was more prevalent to identify mutations or severe genomic irregularities to examine the most probable diseases likely to develop. These methods include risk stratification, predictive modeling, comparative analysis, and genetic risk profiling. While these methods accurately account for single-gene assessments, they do not consider the thousands of other genetic variants that may contribute to disease.

2.4 Limiting Factors

2.4.1 Inconsistent Method Of Collection

Because genetic testing is relatively new, a common standardized procedure for collecting genetic data has not been established. Therefore, insurance companies possess different methods of obtaining data. These differentiated processes include divergent technologies, sampling protocols, and documentation practices, which can affect the genetic results. As a result, the inconsistencies allocate various levels of risk to the same individual causing premium price levels to fluctuate.

2.4.2 Interpretation Challenges

Analyzing genetic data is severely complex as there are thousands of genetic variants that can potentially facilitate the development of a disease. The impact that each genetic variant has on potential diseases in the future is difficult to determine for this reason. As a result of this lack of knowledge, insurance companies are unable to accurately assess an individual's genome and associate it with an equitable premium price.

2.4.3 Non-Genetic Compatibility

While genetic data is quantitative, it is challenging to incorporate it with qualitative variables such as lifestyle, socioeconomic, and behavioral factors. Moreover, genetic data acquired by single-gene analysis is complex to quantitatively define, prohibiting the integration of genetic information with non-genetic variables. For this reason, when creating predictive models it is cumbersome to incorporate a genetic variable with other factors due to them being different data types and methods of measurement. Such incompatibility obstructs risk from being further defined, limiting the proper distribution of premium pricing.

2.4.4 Additional Considerations for PRS Implementation

Real polygenic risk scores face additional challenges including ethnic portability (as most are trained on European populations), the assumption of additive genetic architecture (ignoring epis tasis), and the handling of pleiotropy where single SNPs affect multiple traits. These factors must be considered in real-world implementations.

2.5 Previous Studies

There have been several studies published outlining the benefit of including genetic testing variables into life insurance underwriting decisions. For example, (Lewis et al.) investigated data from the UK Bio-bank in order to research mortality and morbidity outcomes using genetic risk factors. They concluded that genetic predisposition significantly contributes to risk prediction for several different diseases, including cardiovascular disease, stroke, cancer, or diabetes.

Moreover, (Born) explores the effectiveness of traditional underwriting variables and finds that they are insufficient without the addition of genetic testing, while proposing a regulatory approach to accessing genetic information.

As polygenic risk scores are an emerging, unaccredited methodology, many studies have hypothesized the effect that these risk scores have in relation to underwriting decisions. For example, (Karlsson Linnér and Koellinger) demonstrates that polygenic risk scores add enhanced predictive power when integrated with traditional factors, as shown by the different life-span results when utilizing polygenic risk scores.

Additionally, (Maxwell et al. 488–503) examined the potential of polygenic risk scores in predicting common diseases, concluding that polygenic risk scores provide additional risk information, and emphasizing the need for insurers to incorporate genetic data in their risk assessments.

Polygenic risk scores have been found to enhance the accuracy and fairness of risk prediction compared to traditional underwriting variables, as shown by (Lund and Russell 1–7); however, they serve as complementary additions rather than complete replacements.

2.6 Limitations Towards Studies

While there is an extensive amount of data to support that polygenic risk scores are beneficial to insurance companies, their imperfect clinical utility and ethical implications substantiate the need to investigate further. Many of the polygenic risk score models do not include diverse base genetic information, limiting the application of this technology to certain groups. In addition, the accuracy of polygenic risk scores has not been determined as it is a highly complex, emerging technology. Continued research is essential to measure the precise accuracy, and to ensure it is morally acceptable.

III. Data

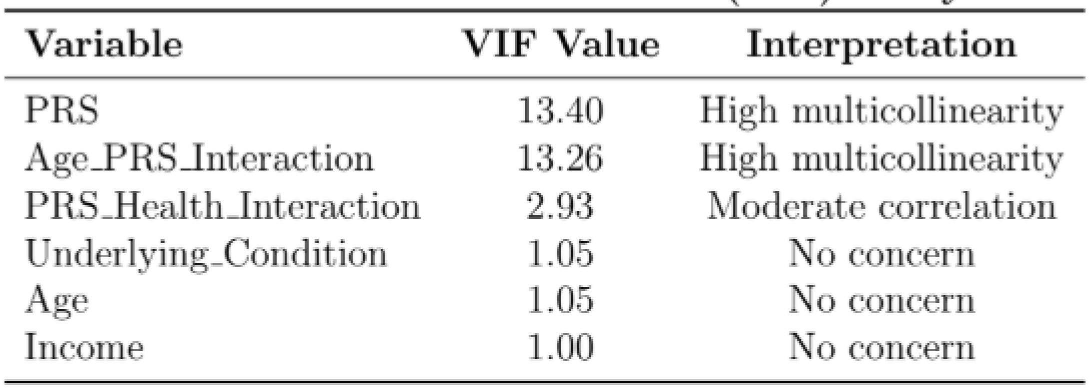
Genetic information is generally difficult to achieve as it provides very sensitive information, such as previous medical diseases, their lifestyle, or unfavorable habits. Because of this, the data that will be provided is simulated, meaning that these individuals are hypothetical. However, these individuals are coded to be realistic so that the data remains useful. It should be noted that the synthetic PRS in this study is a simplified representation and may not capture all complexities of true genetic risk scores.

The manner in which the statistical data was simulated was that 10,000 individuals were created, each of them having a unique combination of a set of variables that would differentiate them from others. The machine learning code used to generate the data and models can be accessed via the following link: https://colab.research.google.com. The variables studied in this analysis have been compiled and are presented (Table 2).

3.1 Multicollinearity Analysis

To address potential concerns about variable relationships, we conducted a multicollinearity analysis. The results revealed some important considerations:

Table 1: **Variance Inflation Factor (VIF) Analysis.**



Variable	VIF Value	Interpretation
PRS	13.40	High multicollinearity
Age_PRS_Interaction	13.26	High multicollinearity
PRS_Health_Interaction	2.93	Moderate correlation
Underlying_Condition	1.05	No concern
Age	1.05	No concern
Income	1.00	No concern

The analysis shows that interaction terms involving PRS exhibit high VIF values, which is expected for interaction terms. The synthetic nature of the data generation process created some artificial correlations, particularly between PRS and family history variables ($r = 0.13$, $p < 0.0001$). These correlations are artifacts of the simulation and should be considered when interpreting results. In real implementations, PRS would be independently derived from genetic data.

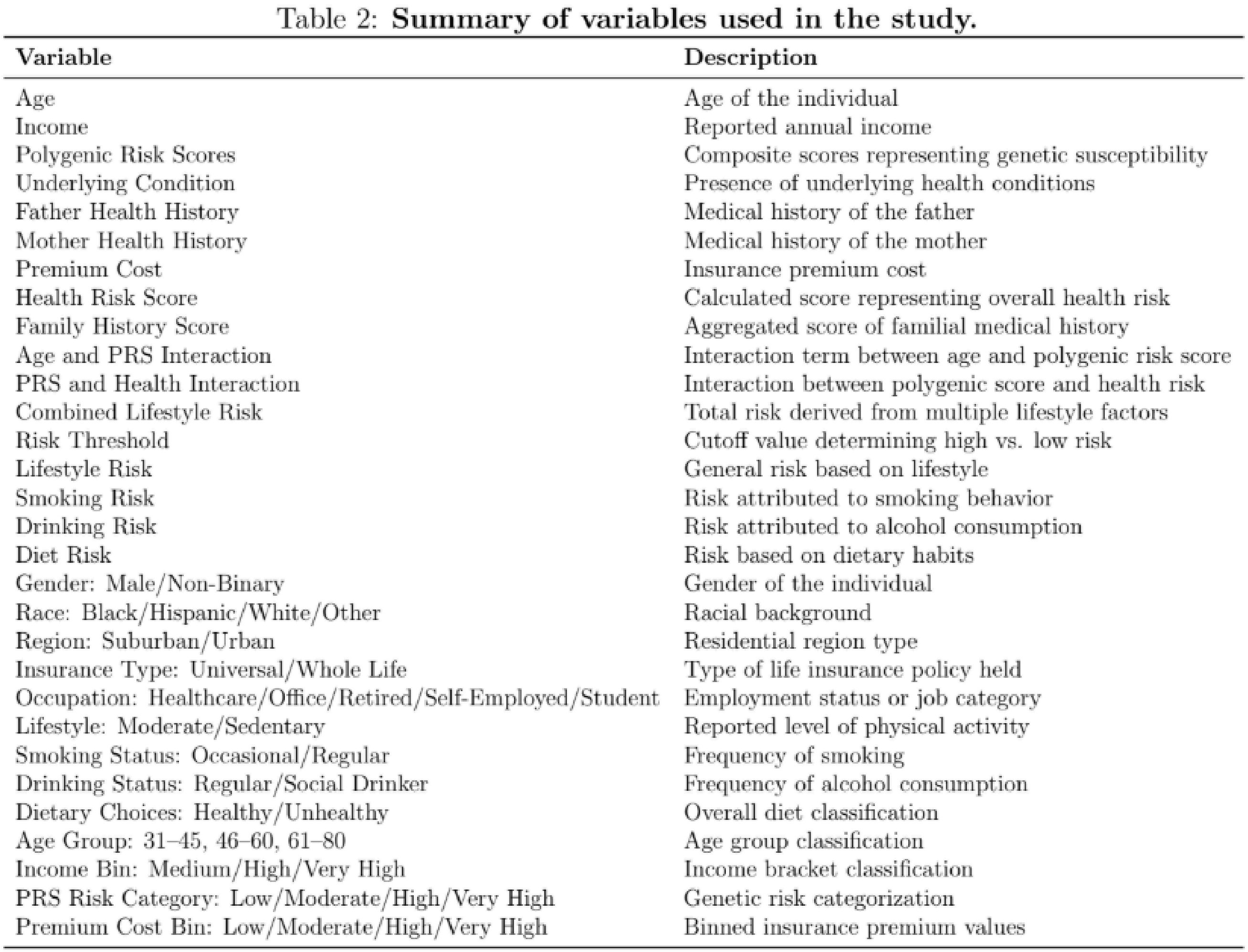


Table 2: **Summary of variables used in the study.**

Variable	Description
Age	Age of the individual
Income	Reported annual income
Polygenic Risk Scores	Composite scores representing genetic susceptibility
Underlying Condition	Presence of underlying health conditions
Father Health History	Medical history of the father
Mother Health History	Medical history of the mother
Premium Cost	Insurance premium cost
Health Risk Score	Calculated score representing overall health risk
Family History Score	Aggregated score of familial medical history
Age and PRS Interaction	Interaction term between age and polygenic risk score
PRS and Health Interaction	Interaction between polygenic score and health risk
Combined Lifestyle Risk	Total risk derived from multiple lifestyle factors
Risk Threshold	Cutoff value determining high vs. low risk
Lifestyle Risk	General risk based on lifestyle
Smoking Risk	Risk attributed to smoking behavior
Drinking Risk	Risk attributed to alcohol consumption
Diet Risk	Risk based on dietary habits
Gender: Male/Non-Binary	Gender of the individual
Race: Black/Hispanic/White/Other	Racial background
Region: Suburban/Urban	Residential region type
Insurance Type: Universal/Whole Life	Type of life insurance policy held
Occupation: Healthcare/Office/Retired/Self-Employed/Student	Employment status or job category
Lifestyle: Moderate/Sedentary	Reported level of physical activity
Smoking Status: Occasional/Regular	Frequency of smoking
Drinking Status: Regular/Social Drinker	Frequency of alcohol consumption
Dietary Choices: Healthy/Unhealthy	Overall diet classification
Age Group: 31–45, 46–60, 61–80	Age group classification
Income Bin: Medium/High/Very High	Income bracket classification
PRS Risk Category: Low/Moderate/High/Very High	Genetic risk categorization
Premium Cost Bin: Low/Moderate/High/Very High	Binned insurance premium values

Numerical probabilities were assigned based on the several options provided per variable with realistic distributions reflecting an average population. On the contrary, to record income, a lognormal distribution was utilized in order to model real-world income distribution. Correlations between variables were introduced to identify relevant relationships to further model accuracy. After individuals' polygenic risk scores were calculated based on their generated characteristics, a reasonable premium cost was assigned, reflecting a binary decision of either acceptance or denial in relation to an insurance company's decision. The premium cost was realistically simulated by multiplying a base premium of $500 by several risk factors reflecting age, genetic risk, lifestyle, health conditions, and family history, as well as some random noise to mimic real-world variability.

The insurance decision distributions of the 10,000 individuals, along with key metrics, are presented to highlight the vital variables investigated in this study. For each key metric, the mean, median, minimum, and maximum values are reported for acceptance and denial, providing information on the relative amounts associated with each decision (Table 3).

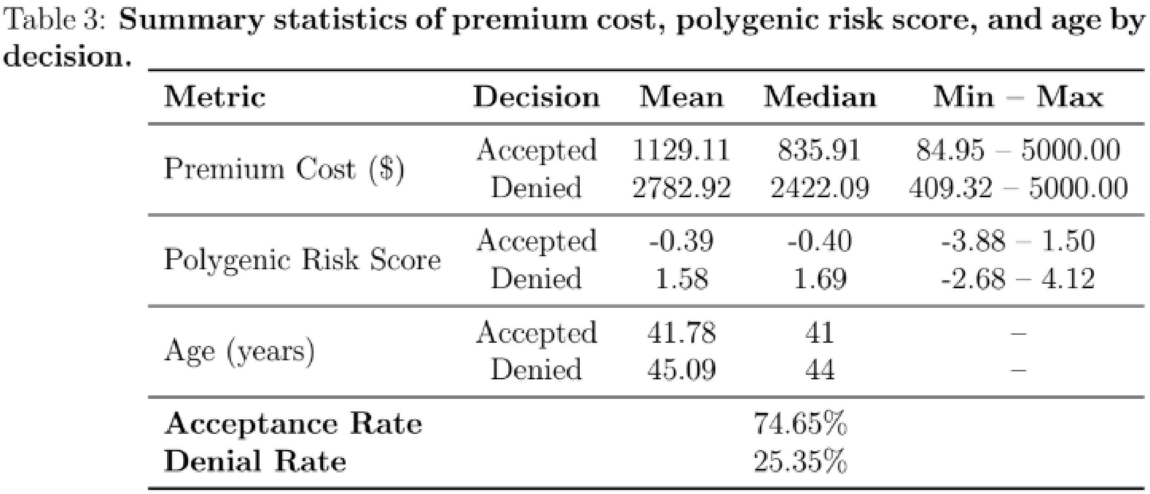


Table 3: **Summary statistics of premium cost, polygenic risk score, and age by insurance decision.**

Metric	Decision	Mean	Median	Min – Max
Premium Cost ($)	Accepted	1129.11	835.91	84.95 – 5000.00
	Denied	2782.92	2422.09	409.32 – 5000.00
Polygenic Risk Score	Accepted	-0.39	-0.40	-3.88 – 1.50
	Denied	1.58	1.69	-2.68 – 4.12
Age (years)	Accepted	41.78	41	–
	Denied	45.09	44	–
Acceptance Rate			74.65%	
Denial Rate			25.35%	

Table 3: Summary statistics of premium cost, polygenic risk score, and age by insurance decision.

In order to improve the performance of the machine learning model, variables were binned, or categorized into discrete ranges to create a simplified set of input options. For example, when grouping age ranges, there is no significant difference between 31-45 years of age with regard to risk. Therefore, they are categorized together to make the data more concise. Binning helps reduce excess variability, making complex data more predictable and easier to interpret.

In addition, one-hot coding was integrated in order to establish numerical values for all variables. While certain variables such as lifestyle, drinking, and smoking can be quantified by a risk score, variables such as an individual's parents health status and an underlying condition are unable to be given a precise score as they are not as measurable. By encoding these unquantifiable values as 1 (present) or 0 (absent), they can be included in the machine learning model and integrated into the study.

IV. Methods

Several methodologies were implemented to analyze the effect of individuals' characteristics and polygenic risk scores on life insurance underwriting decisions.

4.1 Data Splitting and Target Variable

The target variable selected for modeling is the insurance decision, a binary outcome. This is attributable to the fact that insurance decisions can yield solely two results: accepted or denied. The insurance decision is represented by the variable y, while the features—underwriting variables—are represented by x, allowing for the formulation of a predictive relationship between the two.

Moreover, to prevent overfitting, the characteristic of machine learning models to memorize training data rather than generalize underlying patterns, and to obtain an unbiased result, the data set was partitioned into training and testing subsets using an 80/20 split, accordingly. This ratio allows for a perfect balance to enable development of the machine learning model while evaluating model performance on unseen data. As supported by extensive empirical evidence, the 80/20 split is a well-established strategy in predictive modeling. The manner in which the data is split is by stratified sampling, a process to ensure specific strata, or data sets, are represented proportionally by dividing the population into groups based on a particular characteristic, therefore maintaining the original distribution while reducing variances.

4.2 Model Development and Training

Two supervised learning models were developed and trained. These models include a logistic regression model and a random forest model, each displaying disparate advantages. In addition, each model includes a feature importance model, displaying the degree of influence attributed to each variable. However, the development of the models possess identical factors, including 54 engineered features used as inputs. Specifically, different varieties of demographic, clinical, and genomic variables were assessed.

4.2.1 Logistic Regression Model

A logistic regression model is advantageous for many reasons. Known for its simplicity and interpretability, it can highlight underlying patterns with ease. It is generally utilized when the target variable is binary and the relationship is linear.

In this study, a logistic regression model is integrated in order to capture the binary results of insurance decisions. In addition, the model was configured to handle large scale datasets through multinomial setting and saga solver, an optimization algorithm suited for large, sparse datasets due to its efficiency and support for regularization (Chen, Xu, and Liu 1928). To prevent overfitting by developing stronger regularization, a c value of 0.5 is allocated. To eliminate the issue of class imbalance, the weight of each variable was designated as equivalent.

4.2.2 Random Forest Model

A random forest is a machine learning technique that builds multiple decision trees and combines their outputs, offering unique advantages. These include a higher accuracy compared to other models, reduced overfitting, and accurate handling of missing values by considering multiple decision trees. It is generally utilized when addressing complex factors that are non-linear.

In this study, a random forest model was integrated in order to capture non-linear correlations. The model implemented several hyperparameters, including 100 estimators, or core decision trees providing diverse combinations of variables, to enhance the models accuracy and robustness. To prevent overfitting, a maximum depth of 15 layers per tree was implemented. In pursuit of the same goal, the minimum number of samples required to be at a leaf node, or end of a branch is at least 20. In addition, a minimum of 50 samples per leaf node was set to improve model generalization. The square root parameter limits the number of features considered at each split to the square root of the total features, promoting model diversity by using feature subsets rather than all features.

4.3 Evaluation Metrics

We used several metrics to assess the performance of each of the ML models used. Cross validation accuracy helps assess a model's reliability and generalization to unseen data. For all equations referenced below, the abbreviations TP indicates true positive results, TN indicates true negative results, FP indicates false positive results, and FN indicates false negative results. These arise from the errors generated by the models.

Accuracy represents the proportion of correct predictions over total predictions. The formula to calculate accuracy is depicted below.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision demonstrates the amount of positive results that were correct of all predicted positives. A high precision indicates a slight amount of false positive results. It is generally utilized when false positives have serious consequences, and to accumulate more confidence in the amount of true positive results. The formula to calculate precision is displayed.

$$\text{Precision} = \frac{TP}{TP + FP}$$

Similarly, recall indicates the amount of all actual positive cases. Recall is integrated in order to ensure that there is a maximum amount of actual positive cases detected rather than false negatives. A higher recall demonstrates fewer false negatives. The formula to calculate recall is depicted.

$$\text{Recall} = \frac{TP}{TP + FN}$$

Due to inaccuracies and probability of deceptiveness, F1 score is integrated as a key metric by balancing precision and recall. It is beneficial when addressing unbalanced data sets. The formula to calculate F1 score is shown below.

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

In addition, the ROC AUC curve is used to assess the model's ability to distinguish between positive and negative outcomes. ROC AUC is implemented when addressing binary decisions. An ROC AUC of above 0.8 or 80% is considered excellent. The formula is displayed below. The N+ and N- represent the number of actual positive and actual negative cases, and the summations are displayed to find the area under the curve. Furthermore, the s_i and s_j variables are part of an indicator function utilized for counting how many times a positive instance is ranked above a

$$\text{AUC} = \frac{1}{N_+ N_-} \sum_{i=1}^{N_+} \sum_{j=1}^{N_-} \mathbf{1}\left(s_i > s_j\right)$$

In order to provide not only a per-class view, several alternate calculation methods were incorporated to convey the overall synthesized results collected from both accepted and denied insurance decisions.

Macro Average simply calculates the average between the binary per-class metrics by considering all classes equal. Consequently, bias towards common classes is omitted, creating a more balanced and comprehensive view of the model's effectiveness. The formula for calculating macro average is exhibited. N is the number of classes, and $M_i$is a metric value of either precision, recall, or F1 Score.

$$\text{Macro Average} = \frac{1}{N}\sum_{i=1}^{N} M_i$$

Similarly to macro average, weighted average calculates the average between the binary per class metrics. Nevertheless, weighted average factors in support, or the number of actual instances, in order to weigh classes by different means. By weighing classes it provides a more realistic overall performance. The formula is demonstrated below.

$$\text{Weighted Average} = \frac{\sum_{i=1}^{N} (\text{Support}_i \times M_i)}{\sum_{i=1}^{N} \text{Support}_i}$$

V. Results and Discussions

In this section, we present our results on the analysis of the machine learning model using logistic regression and random forest models. In addition, the confusion matrices constructed from each model are analyzed, as well as each of their feature relevance.

5.1 Model Performance

The predictive performance of the models was assessed using various evaluation metrics to understand their relative strengths and weaknesses. This comparison provides insight into how well each model distinguishes between accepted and denied cases (Table 4).

Table 4: Comparison of Logistic Regression and Random Forest model performance.

Metric	Logistic Regression	Random Forest
Cross-Validation Accuracy Accuracy F1 Score ROC AUC	0.7301 (±0.0070) 0.7380 0.6048 0.8109	0.9748 (±0.0027) 0.9800 0.9617 0.9982
Classification Report (Class: Accepted) Precision Recall F1-Score Support	0.91 0.72 0.80 1493	1.00 0.98 0.99 1493
Classification Report (Class: Denied) Precision Recall F1-Score Support	0.49 0.79 0.60 507	0.93 0.99 0.96 507
Overall Metrics Accuracy Macro Average F1 Weighted Average F1	0.74 0.70 0.75	0.98 0.97 0.98

5.1.1 Summary Level Metrics

While the logistic regression model includes cross validation accuracy, accuracy, F1 score, and ROC AUC of all below 0.82, the random forest model includes all of these metrics above 0.96. This demonstrates that the random forest model is significantly more effective at distinguishing between accepted and denied cases due to its higher complexity in analyzing data patterns. However, the exceptional performance should be interpreted with caution given the multicollinearity identified in our analysis.

5.1.2 Classification Reports

In addition, when analyzing precision, recall, and F1 score for both accepted and denied cases, the random forest model significantly outperforms the logistic regression model. Whereas the random forest model had a collective of above a 0.93, with also having a perfect precision for the accepted decision of 1.0, the logistic regression model had a collective of below 0.81 with the exception of the precision for the accepted decision (0.91). Both models were more accurate in regards to accepted decisions rather than denial decisions, although the random forest model is significantly more proficient in distinguishing between false positives and false negatives.

5.1.3 Overall Metrics

When comparing the means of the accepted and denied classification metrics using accuracy, macro average, and weighted average, it can be determined further that the random forest model is superior compared to the logistic regression model. For example, the overall metrics of the logistic regression were all under 0.76, while the overall metrics of the random forest model is more than 0.96. This discrepancy indicates the random forest model's superiority in constructing reliable predictions without bias by accounting for class distribution.

In every aspect the random forest model demonstrated enhanced predictive performance compared to the logistic regression model. While both models possess their own advantages, in this study the superior model for accuracy and classification is the random forest model due to its ability to analyze complex data with non-linear relationships. The multicollinearity analysis suggests that some of this performance gain may be due to the model exploiting correlations in the synthetic data.

5.2 Confusion Matrices

The confusion matrices displaying the results of insurance decisions generated by the logistic regression and random forest models are presented (Figure 1).

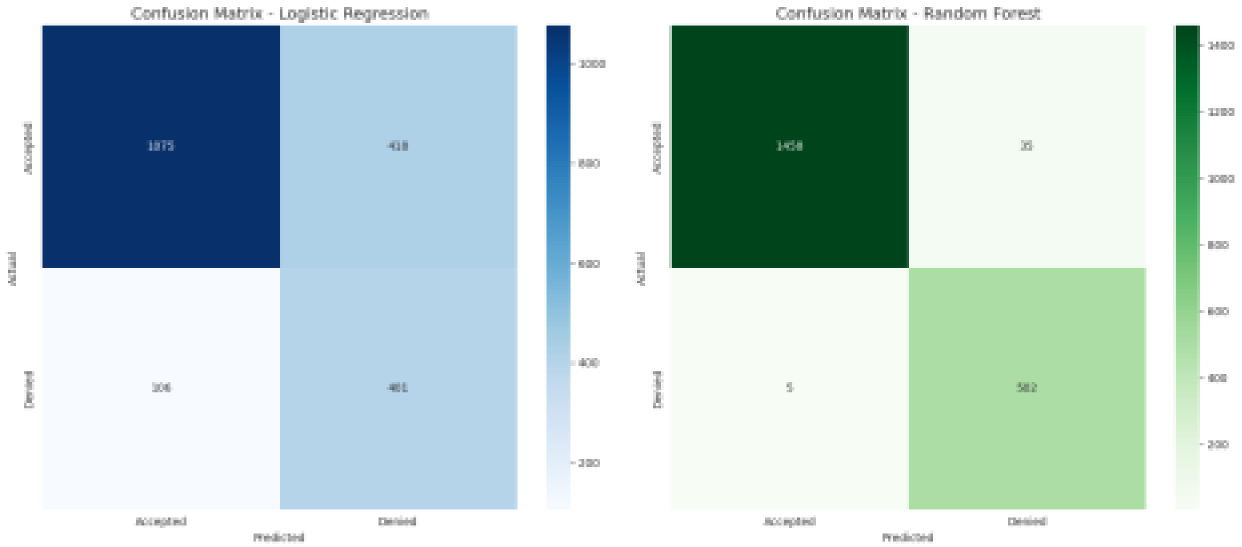


Figure 1: Confusion matrices for the classification models. These heatmap-style matrices display the counts of true positives, true negatives, false positives, and false negatives for each model, helping to visualize their performance in correctly classifying accepted and denied cases, as well as the types of prediction errors made.

As shown in the matrices, the horizontal portion of the graph indicates the predicted accepted or denied cases, while the vertical portion predicts the actual accepted or denied cases. A heat map is included in order to easily identify which predictions compared to actual results occurred the most. The matrices highlight the four possible classification outcomes: true positive, true negative, false positive, false negative.

While both confusion matrices performed adequately, the random forest model was more accurate in its predictions. For true positive and true negative cases, the random forest model produced 1,458 and 502 instances, respectively, whereas the logistic regression model produced 1,075 and 401 instances. This demonstrates the random forest model's superiority in accuracy as it identified more correct decisions than did the logistic regression model. For false positive and false negative cases, the random forest model yielded 5 and 35 instances, respectively, while the logistic regression model yielded 106 and 418 instances. This exhibits the dominance of the random forest model as it identified fewer incorrect decisions.

5.3 Feature Importance

The feature importance charts for the random forest and logistic regression models are displayed and were evaluated to identify the underwriting variables most influential in predicting outcomes (Figure 2).

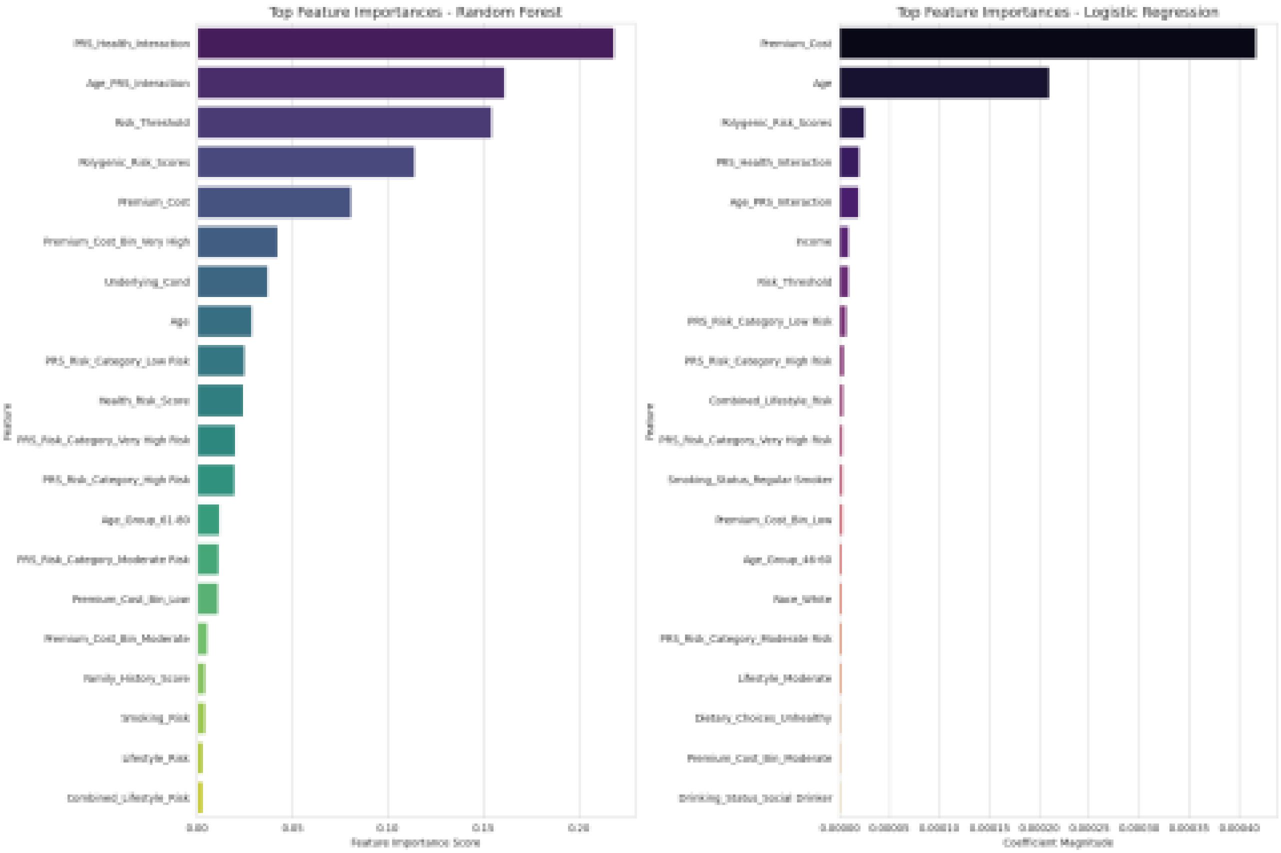


Figure 2: Feature importance by model. This figure illustrates the contribution of each variable to the models' prediction results, highlighting the most influential features that impact insurance decisions. Understanding these key drivers helps interpret how the models evaluate risk. Note that the high importance of interaction terms may be influenced by the multicollinearity identified in the data.

As shown by the feature importance results, there is a clear difference between the variable significance of the logistic regression model and the random forest model. The logistic regression model mainly places substantial significance on two variables: age and premium cost. In contrast, the random forest model exhibits more diverse feature importance through several variables with polygenic risk score interactions and premium costs being among the most influential. These deviations help explain the difference in the accuracy among the two models. However, the high importance of interaction terms (Age PRS Interaction, PRS Health Interaction) should be interpreted carefully given their high VIF values.

By spreading out the influence of the underwriting decision variables and putting more weight towards polygenic risk score variables, the random forest model is able to reduce the susceptibility to bias. Therefore, the random forest model is more accurate than the logistic regression model due to the significance given to polygenic risk scores, though this may be partially due to the synthetic nature of the data.

In addition, the different variables that were weighed significantly impacted the accuracy of the logistic regression and random forest models. The results of the random forest model indicate it as the more accurate model in terms of indicating correct predictions. Therefore, we can consider the weight of the variables of the random forest model as superior indicators than the ones used for logistic regression. In the random forest model, polygenic risk score interactions had a greater importance, in addition to premium cost, underlying condition, age, and health risk score. This indicates that by utilizing polygenic risk scores in determining the insurance decision of an individual, it significantly increases the accuracy of risk prediction, therefore providing a more accurate risk-assessment of an individual. In the logistic regression model, age and premium cost are mainly the only two variables that are significantly weighed. By mainly considering these demographic factors it over-simplifies the risk-level of individuals, leading to limited accuracy in predictive insurance decisions.

5.4 Limitations and Considerations

While our results show promising improvements in model performance with PRS inclusion, several limitations must be acknowledged:

- The synthetic nature of the PRS may not fully capture the complexity of real genetic risk scores, which require integration of thousands to millions of SNPs
- Multicollinearity analysis revealed high VIF values for PRS-related interaction terms, suggesting potential redundancy
- The artificial correlations between PRS and family history variables are artifacts of the data generation process
- Real PRS implementation would face challenges including ethnic portability, as most are trained on European populations
- The assumption of additive genetic architecture may not reflect biological reality

These limitations suggest that while our study demonstrates the potential value of genetic information in insurance underwriting, real-world implementation would require addressing these methodological challenges.

VI. Conclusions

In this study, we explored the effect of polygenic risk scores and other variables in life insurance underwriting decisions using machine learning, such as logistic regression and random forest models. Using the models, we constructed confusion matrices and feature importance plots to illustrate their predictive performance.

Our analysis indicated that overall, the random forest model was clearly the superior model when analyzing the predictive accuracy of insurance decisions due to its ability to analyze complex, non-linear data. In addition, polygenic risk scores and their interactions significantly enhanced the predictive accuracy of insurance decisions as when they were weighted more, the predictive accuracy was much greater compared to when weighing mainly demographic and less complex variables, as shown by the comparison between the random forest model and the logistic regression model. However, the multicollinearity analysis revealed that some of these improvements may be due to artificial correlations in the synthetic data rather than genuine predictive value.

While polygenic risk scores substantially aid in improving predictive accuracy of insurance decisions, other variables remain equally important in determining life insurance underwriting decisions. In the feature importance analysis, the logistic regression significantly considers age and premium costs, and the random forest model significantly considers premium cost, underlying condition, age, and health risk score. These traditional underwriting variables are persisting to be used in conjunction with genetic information, as demonstrated by the feature importance chart and the prediction accuracy of the models.

These results emphasize the potential importance of including genetic variables into life insurance underwriting decisions, while acknowledging the challenges of implementation. Several factors—among the most important being polygenic risk score and their interactions, premium cost, underlying condition, age, and health risk score—substantially impact the effectiveness of risk-prediction models. The random forest model demonstrates the importance of these factors by producing prominent results in predictive accuracy, though the high VIF values suggest caution in interpretation.

In conclusion, integrating genetic testing, specifically polygenic risk scores, with various traditional variables shows promise for improving predictive accuracy in underwriting methods. How ever, real-world implementation would require addressing several challenges:

• Development of true PRS from GWAS studies rather than synthetic
approximations • Addressing ethnic bias and ensuring fairness across populations

• Resolving multicollinearity issues between genetic and traditional
variables

• Establishing regulatory frameworks to prevent genetic discrimination

• Ensuring transparency and explainability of models using genetic information

By addressing these challenges and applying properly validated genetic variables to risk-prediction models, insurance companies could potentially more accurately detect the risk-level of individuals, leading to reduced adverse selection and increased sustainability of insurance companies, while maintaining fairness and ethical standards.

References

Low, Lisa, et al. “Genetic Discrimination in Life Insurance: Empirical Evidence from a Cross Sectional Survey of Genetic Support Groups in the United Kingdom.” *BMJ: British Medical Journal*, vol. 317, no. 7173, 1998, pp. 1632–1635. https://doi.org/10.1136/bmj.317.7173.1632.

Lewis, Cathryn, et al. *Epidemiological and Genetic Prediction of Major Morbidity and All Cause Mortality*.UKBiobank,2021,www.ukbiobank.ac.uk/enable-your-research/approved-research/investigate-the-importance-of-genetic-data-over-and-above-detailed-epidemiological-data-in-multivariable-prognostic-models-that-predict-risk-of-i-major-morbidity-ii-all-cause-mortality. Accessed June 2025.

Born, Patricia. “Genetic Testing in Underwriting: Implications for Life Insurance Markets.” *Journal of Insurance Regulation*, vol. 38, no. 5, 2019.

Karlsson Linnér, Richard, and Philipp D. Koellinger. “Genetic Risk Scores in Life Insurance Underwriting.” *Journal of Health Economics*, vol. 81, 2022, p. 102556.

Maxwell, Jessye M., et al. “Multifactorial Disorders and Polygenic Risk Scores: Predicting Common Diseases and the Possibility of Adverse Selection in Life and Protection Insurance.” *Annals of Actuarial Science*, vol. 15, no. 3, 2021, pp. 488–503.

Lund, Heather, and Richard Russell. “Polygenic Risk Scores: A Useful Tool in Our Risk Prediction Toolkit?” *Japanese Journal of Insurance Medicine*, vol. 117, no. 1, 2019, pp. 1–7.

Chen, Zhihua, Xuchen Xu, and Hongbo Liu. “The Successive Approximation Genetic Algorithm (SAGA) for Optimization Problems with Single Constraint.” *Mathematics*, vol. 11, no. 8, 2023, p. 1928.