AMERICAN ACADEMIC PUBLISHER

INTERNATIONAL JOURNAL OF NETWORKS AND SECURITY (ISSN: 2693-387X)

Volume 05, Issue 02, 2025, pages 01-29
Published Date: - 07-07-2025
Doi: -https://doi.org/10.55640/ijns-05-02-01

Predictive Risk Modeling in P&C Insurance Using Guidewire DataHub and Power BI Embedded Analytics

Kawaljeet Singh Chadha
University of the Cumberlands, Williamsburg, KY, USA

ABSTRACT

P&C insurers are increasingly pressured to identify and effectively predict risk. While traditional methods, such as actuarial models and manual assessments, are effective for identifying patterns in large-scale policy and claims data, they struggle to capture complex patterns, like resistance curves. This paper examines how predictive risk modelling can be implemented in practice using Guidewire DataHub and Power BI Embedded Analytics. Power BI is used for interactive visualization and real-time decision support, whereas Guidewire Data Hub is utilized as a centralized platform for storing and managing structured insurance data. It utilized structured data from claim history, underwriting attributes, policy details, and customer profiles to build a predictive model. Machine learning algorithms, such as Random Forest and Logistic Regression, were then applied to classify policyholders as High, Medium, or Low risk after preprocessing and feature selection. Standard metrics (accuracy, precision, recall, ROC-AUC) were used to evaluate model performance. The Random Forest classifier achieves an accuracy of 84% and identifies high-risk profiles most effectively. It then integrated these predictions with Power BI dashboards, allowing underwriters and analysts to explore risk at both the individual and portfolio levels. The study illustrates how building data platforms that integrate machine learning and embedded analytics facilitates more innovative underwriting, fraud detection and pricing. In a competitive, data-driven insurance environment, the ability to turn raw insurance data into actionable insights provides significant operational and strategic value.

KEYWORDS

Predictive *Risk Modeling, Property and Casualty Insurance, Guidewire DataHub, Machine Learning in Insurance, Power BI Embedded Analytics*

1. INTRODUCTION

Risk evaluation is the backbone for all policy decisions, pricing strategies and operating mechanisms in the insurance industry. Insurers within the Property and Casualty (P&C) segment grapple with a broad range of risk types, from automobile collisions to property damage, natural disasters, and fraudulent claims. To effectively manage these risks, one must understand how likely future events are and their potential impact. Historical data analysis, underwriters' professional judgment and actuarial models have traditionally been the basis for assessing risk. While these approaches remain useful, they are insufficient to address the complexity, scale, and velocity of modern insurance data.

With readily available digital data and the attraction to cloud-based platforms, the entire insurance process has

undergone a massive significant transformation. As insurers shift to data centric operations, attention is now focused on moving from static reporting to dynamic, real-time analytics. It allows a proactive and more informed decision process. However, data is meaningful only subsequently, and yet raw data alone is of little value; it is what is made of it that counts. The capability to predict meaningfully has become a critical gap that predictive modelling helps address. Predictive models learn from historical patterns to estimate the probability of future events occurring, such as future policyholder claims, premium payment lapses, or potential fraud. Using these insights, underwriting accuracy can be improved, pricing structures refined, claims triage can be enhanced, and operational losses can be reduced. However, with the advent of modern technology solutions, it is now possible to integrate predictive analytics directly into business processes.

One such platform is Guidewire DataHub, which is built specifically for the insurance industry. This unifies the repository of data and brings it to a single place, capturing and organizing management, billing systems, and customer touchpoints, reducing the need for researchers to gain access to clean and consolidated data necessary for reliable analytics. When combined with other business intelligence tools, such as Microsoft Power BI, this data serves as a basis for building interactive dashboards, visual risk monitoring, and enterprise reporting.

The construct of the present study aims to develop a predictive risk classification model based on structured data from a simulated Guidewire DataHub environment. The idea behind the project is to categorize policyholders into distinct risk categories (e.g., low, medium, high risk) by examining indicators related to their behavior and historical data. The model is then trained using supervised machine learning algorithms aided by features extracted from policy, claim and customer datasets. The model is then trained and validated, and the outputs are then integrated into an embedded Power BI dashboard. Insurance professionals utilize this dashboard to visualize risk scores, claim trends, and feature-level insights in real time, thereby eliminating the need for technical teams to be involved in the process. To offer a scalable and practical solution for improving risk assessment in P&C insurance, this approach combines advanced analytics, data warehousing, and interactive visualization. Greater accuracy and visibility in classifying risk enhance operational efficiency and support strategic decision-making, all of which offer a competitive advantage in a data-driven marketplace.

2. LITERATURE REVIEW

This section reviews current research and professional practices in predictive risk modelling for P&C insurance (4). It covers developments in data platforms, analytics tools, and modelling techniques that enhance underwriting, pricing, and fraud detection.

2.1 Insurance Predictive Risk Modelling

In modern insurance analytics, predictive modelling is a major player. Primarily, these models are used to estimate claim likelihood, detect fraud, and provide a basis for pricing. Logistic regression, decision trees, and ensemble models are widely used supervised machine learning algorithms for classifying risk using historical data. These models can process more variables and identify more complex patterns more quickly than traditional actuarial techniques while still acknowledging their limitations when compared with manual analysis. Risk assessments increasingly utilize behavioral and transactional variables, such as previous claims, payment delays, and changes in policy coverage. These variables provide insight into behavior and risk. The availability of larger and more detailed datasets has enabled the achievement of better model performance and more precise segmentation.

As shown in the figure below, predictive analytics in insurance supports a wide range of objectives, including improved risk assessment, fraud detection, cost reduction, enhanced customer experience, and regulatory

compliance.

Figure 1: predictive-analytics-in-insurance-industry

2.2 Risk Modeling Driven by Data Platforms

Insurers can manage and structure information from multiple departments using modern data platforms. New systems, such as Guidewire DataHub, provide a single location for consolidating policy, claims, billing, customer service data, and more. This centralization improves consistency across applications and simplifies analytics by eliminating data duplication. By creating a unified data environment, these platforms allow risk models to be built using clean, reliable inputs. Moreover, structured and well-governed data makes it easier to comply with regulatory standards and internal audit protocols. As noted by Goel, building resilient systems—whether in supply chain or data management—depends on reducing fragmentation and enabling robust foundational infrastructures (14).

2.3 Embedded analytics tools

This provides non-technical users, such as underwriters and claims managers, with the ability to interact with model results in real-time, using embedded analytics tools like Microsoft Power BI. Organizations can build dashboards that depict essential insights, including claim trends, risk levels, and possible fraud indicators, using these tools. Dashboards offer visual summaries that enable informed decisions and eliminate the need for static reports (12). Embedded analytics combines real-time data feeds with user-friendly visuals to increase transparency and provide alignment of decisions between departments. The added value lies in the ability to drill down into individual policy details or gain insight into trends at a portfolio level.

2.4 Machine Learning Techniques for P&C Risk Assessment

In the insurance sector, a wide range of machine learning models is used. Ensemble methods, such as Random Forest and XGBoost, are renowned for their high accuracy and their ability to work effectively with various types of variables. Furthermore, they can determine which variables contribute the most to risk predictions (supporting the explainability). Insurance has also been tested with advanced models, including neural networks, to determine complex patterns such as fraud. These models, however, are less transparent and more difficult to explain to

business users or regulators. Therefore, in many occasions, environments where transparency is essential, simpler models with explainable logic are preferred. SHAP, an explainable AI tool, is used to explain complex models and provide detailed feedback on individual predictions. These serve to maintain faith and ensure that, if necessary, an export format for audits and reviews of model outputs remains possible.

As shown in the figure below, machine learning supports various underwriting tasks including creditworthiness assessment, fraud detection, insurance risk assessment, pricing strategies, and portfolio management.

Figure 2: underwriting-machine-learning-models

2.5 Gaps in the Existing Literature

Predictive modelling is a proven area; however, some unexplored areas within this field also have significant gaps. One notable gap is the lack of studies documenting how data platforms, modelling toolkits, and visualization dashboards can be integrated cohesively into a unified system. The existing research is mainly focused on modelling or data, but not on the combination of both. Ensuring consistency and real-time responsiveness is essential to operationalizing analytics-driven systems, which parallels the need for integration in insurance predictive modelling (10). Insurers also lack guidance on the use of predictive models in operational live environments, where they must make their results accessible and understandable. This paper fills these gaps by presenting an end-to-end, practical approach to risk modelling built on Guidewire DataHub and Power BI, aligning with the call for scalable, performance-driven solutions in complex data infrastructures (11).

3. Theoretical Framework

3.1 Risk Modeling Principles in P&C Insurance

Property and Casualty (P&C) insurance risk modelling is used to estimate the likelihood and magnitude of future losses that may occur due to a claim, policyholder behavior, or an external event (19). Vehicle collisions, property damage, liability exposure and insurance fraud are common types of risk. These risks have traditionally been handled by insurers using actuarial techniques. However, such methods require stable, linear relationships and are mostly limited to low-dimensional, low-dimensional data. A predictive modelling approach provides a more

adaptive solution, leveraging historical patterns, behavioral indicators, and real-time data inputs to forecast. These models are key for accurate underwriting, customer segmentation and proactive claims management.

As illustrated in the figure below, the P&C insurance sector is rapidly evolving, driven by multiple industry trends. These trends—including digital transformation, predictive analytics, and customer-centric innovations—are reshaping traditional risk modeling strategies and highlighting the need for adaptable, data-driven approaches.

Figure 3: Property & Casualty Insurance Industry Trends

3.2 Predictive Analytics and CRISP-DM Process.

The Cross Industry Standard Process for Data Mining (CRISP-DM) provides a structured method for developing and applying predictive models (27). The first step of this process involves understanding the business problem (in the insurance case, this might mean improving the accuracy of underwriting or reducing the fraudulent claims rate). Once the objective is defined, the next step is to understand and explore the dataset, then clean, transform and prepare the data for modelling. Then, they select the algorithms needed to build the models and evaluate them through metrics such as accuracy and error rate. The final stage is to provide a way for the models to be deployed in a system that decision-makers can access and use. The CRISP-DM framework ensures that the modelling process is aligned with business needs and that this process is repeatable within a disciplined workflow.

3.3 Feature engineering and dimensionality reduction.

Insurance data is often diverse and voluminous, consisting of structured information such as policy attributes, customer demographics, vehicle specifications, and historical claim records. The choice of algorithm is not the sole determinant of modelling quality; the quality of the input features also matters. Feature engineering involves selecting, transforming, or creating variables that serve as proper signals for the prediction task. Encoding categorical variables, normalizing numerical fields, and handling missing data may be involved in this process. Dimensionality reduction techniques are employed to enhance model performance and interpretability further. Using correlation-based feature selection, it selects and stores only variables that have a strong statistical

relationship with the target variable, eliminating others that include noise or redundancy. Principal component analysis is one 'feature reduction' algorithm that simplifies the dataset by transforming the original features into a smaller set of derived components that capture the most variance in the data (1). Furthermore, these methods can streamline the modelling process and reduce the risk of overfitting.

3.4 Risk Classification with Machine Learning

Several widely used supervised machine learning algorithms are employed in insurance risk classification. Logistic regression is a linear, logically interpretable model for predicting an outcome as a function of multiple inputs. Decision trees and random forests are flexible models that can convert numerical data with categorical data and generate easily understandable decision rules. Gradient boosting and other methods using ensemble combine several base models into a structured sequence, leading to higher prediction accuracy. Neural networks and deep learning methods can model complex, non-linear relationships, but often in less transparent and more challenging environments for regulatory applications. Both algorithms and their characteristics are a function of the data (including its size and dimensionality), the desire for interpretability and the operational environment. Error metrics, such as mean absolute error (MAE), root mean squared error (RMSE), and area under the receiver operating characteristic curve (AUC), are used to assess models and ensure they are both accurate and reliable in their predictions.

3.5 Business Intelligence Integration & Decision Support

Predictive risk modelling is the ultimate goal, helping to support decision-making within the insurance organization (38). Model outputs must be presented in a clear and accessible format to be actionable. Interactive dashboards created with Power BI Embedded are business intelligence tools that visually render model results, allowing users to explore them. Directly connected to live data sources from Guidewire DataHub, these dashboards provide current insights, including customer risk scores, claim frequency trends, and regional risk concentrations. These insights can be used by users (such as underwriters, analysts, and claims managers) to select priority cases, adjust prices or flag suspicious activity. Using a visual and data-driven approach speeds up and helps make more informed and transparent decisions across sections. Analytics are integrated into daily operations, ensuring predictive models deliver meaningful business value rather than remaining closed to technical teams, similar to how integrated systems in healthcare and security pipelines enhance operational responsiveness (23,34).

4. System Architecture Overview

4.1 Overview of Architecture Design

Predictive risk modelling is designed to be supported by an architecture that enables seamless integration between core insurance data systems, machine learning environments, and business intelligence tools. The five major components include data extraction from Guidewire DataHub, data transformation and staging, centralized storage, training machine learning and embedded analytics with Power BI. The modular design of the system ensures that it can scale, be maintained, and be made accessible to data science and business user teams (36,37).

4.2 ETL Pipeline from Guidewire Data Hub

Data extraction begins with scheduled jobs or API-based connectors that extract structured data from Guidewire DataHub. Policy records, claim transactions, billing activity, and customer profile information are all extracted datasets. The transformation layer processes these datasets by imputing missing values, encoding categorical

variables and deriving new ones. The transformed data are placed in a staging layer for use in validation and quality control. Standardizing data in this step ensures that data aligns with enterprise data governance policies. The image below illustrates a standard ETL (Extract, Transform, Load) architecture. Data from various sources—including RDBMS, API endpoints, cloud services, and flat files—is first extracted and moved to a staging area.

The image below illustrates a standard ETL (Extract, Transform, and Load) architecture. Data from various sources—including RDBMS, API endpoints, cloud services, and flat files—is first extracted and moved to a staging area.

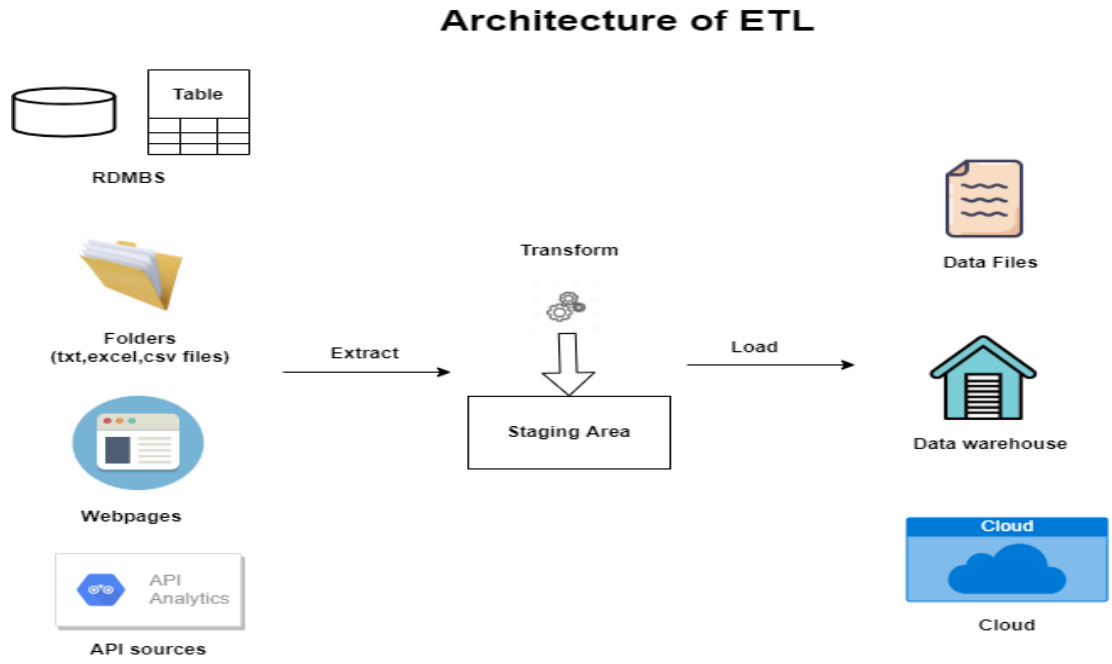


Figure 4: ETL data pipelines

4.3. Data Lake and Model Training Environment

Once pre-processed, data is loaded into a secure data lake environment. This is the central place where raw and processed insurance data sit (35). It is designed to handle massive volumes of structured records and supports both batch and incremental updates. Datasets are pulled from the data lake to the machine learning workspace, which is populated with Python-based tools (Pandas, sci-kit-learn, XGBoost) specific to the case study at hand. Models are trained for this environment using supervised learning algorithms ie. Random Forest, Logistic Regression.

4.4 Reporting Layer and Power BI Embedded Integration

Outputs of the model, such as risk scores, class labels, and feature importance's, are pushed to a reporting database or published as API endpoints once the model is trained. The outputs of these data pipelines are then consumed by Power BI Embedded, which creates interactive dashboards for business users to view. Key risk indicators are presented at the policy, customer, and portfolio levels through dashboards. Heat maps, claim trend lines, classification summaries, and filterable reporting by geography, policy type, or time frame are visual elements. Embedded within internal web-based applications or portals, these interfaces provide near real-time access to insights for users, including risk analysts and underwriters, without requiring them to leave one platform (21, 22).

4.5. Data Governance, Security and Compliance

Data governance measures are firmly integrated with the architecture. Access control is implemented via role-

based permissions, thereby limiting a user's interaction to only what they should be interacting with based on their role. Modelling and dashboard reporting mask personally identifiable information to protect our customers' privacy. All activities related to data processing and model prediction are logged for traceability and auditability. The architecture adheres to industry standards, such as GDPR and local insurance regulations, regarding data handling, which means that predictive models can be deployed in an operationally and legally compliant manner.

5. Dataset Description

5.1 Overview of Data Source

The structured records used in this predictive risk modelling study consist of simulated Guidewire DataHub records that conform to the structure of actual P&C insurance data systems (30). It contains anonymized and synthetic records that represent the kind of data insurance companies collect during a policy lifecycle. The data spans multiple years and includes records about customer profiles, policy issuance, endorsements, claims events, payments, and renewals. This precisely defines the structure of the data so that it not only represents real-world insurance operations but also provides sufficient details to enable meaningful risk classification and prediction, supporting the need for fault-tolerant, context-aware data environments as emphasized in modern event-driven architectures (5, 6).

The visual below outlines the typical cycle followed in predictive analytics. It begins with data collection, followed by model selection, training, deployment, and continuous monitoring. This process mirrors the workflow used in this study to build, test, and integrate the insurance risk prediction model based on Guidewire DataHub data.

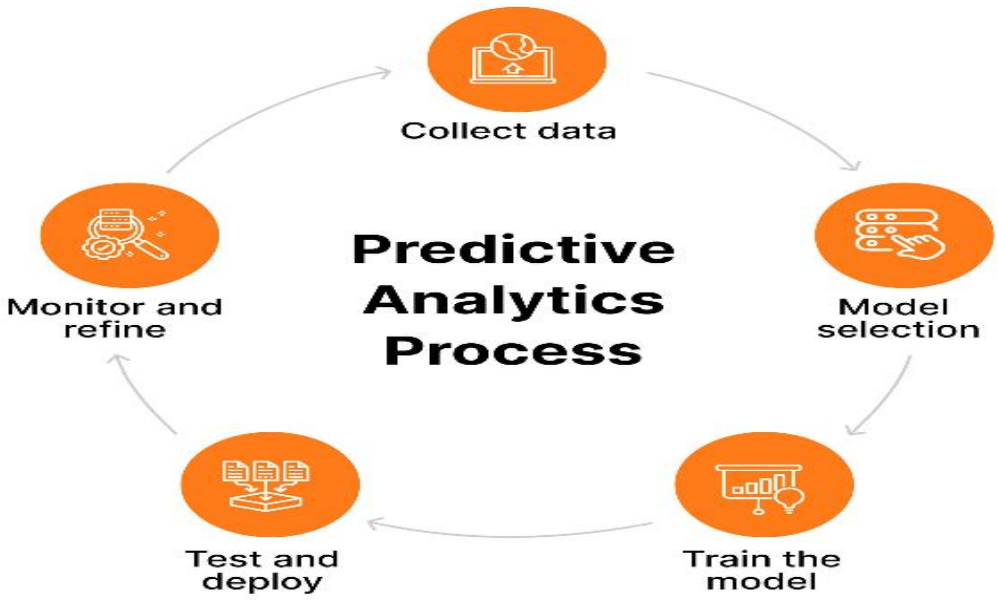


Figure 5: automation-and-ai-role-in-enhancing-predictive-analytics-for-marketers

5.2 The volume and structure of the data

It includes about 58,000 policy records over 120 attributes. The values are a mix of categorical, numerical, and date/time (7). Customer age, policy type, coverage limits, vehicle details, number of claims filed, claim amounts, geographic location, payment patterns, and policy tenure are some of the key fields. The dataset consists of

transactional records, such as claims and billing events, which are linked together by a unique customer ID and policy ID; each record in the dataset maps to an individual policyholder. The tabular format (CSV and Parquet) data was easy to integrate with external modelling tools and visualization platforms as it was exported from a sandbox version of Guidewire DataHub.

Various types of data attributes used in the analysis, including both numerical and categorical variables, are summarized in the Table below.

Table 1: ***Data Types and Examples***

Data Attribute	**Type**	**Example Value**
Customer Age	Numerical	42
Policy Type	Categorical	Comprehensive
Vehicle Value	Numerical	25,000
Region	Categorical	Northeast
Number of Claims Filed	Numerical	2
Claim Amount	Numerical	6,300
Payment Method	Categorical	Auto-Debit

5.3 Data Quality and Data Preprocessing

The dataset was assessed for completeness, consistency, and accuracy before being analyzed. Much of the data included in these fields (with optional fields such as secondary driver information or claim cause details) was either missing or inconsistent. The multiple imputation technique, using chained equations, was employed to handle missing data. This fills in missing values by modelling each field as a function of the other fields, making it more robust to recovery than simple mean or median substitution (32). In fields like total claim amount and policy premium, outliers were also identified. To mitigate the impact of extreme but rare cases, these values were capped at the 95th percentile. To prepare categorical fields such as region, vehicle type, and policy channel for machine learning algorithms, these were encoded via one-hot encoding. To avoid model training being skewed by attributes on differing scales (such as age and claim amount), numerical variables were normalized. Derived features were also created to augment the predictive power of the model. Some examples of these fields include claim frequency per policy year, time since the last claim, payment delinquency score, and loss ratio (historical). Architecting these engineered features is crucial in distinguishing risk profiles that would otherwise appear identical based on the raw data.

The key variables incorporated into the predictive model, along with their respective types and importance levels, are summarized in the Table below.

Table 2: Key Variables Used in the Model

Variable Name	Type	Role in Model
Claim Frequency	Numerical	High Importance
Loss Ratio	Numerical	High Importance
Payment Delinquency Score	Numerical	Medium Importance
Time Since Last Claim	Numerical	Medium Importance
Policy Tenure	Numerical	Medium Importance
Vehicle Type	Categorical	Low Importance
Geographic Region	Categorical	Low Importance

5.4 Definition of the target variable

This dataset has a categorical label as its response variable, meaning they are the risk class of each policyholder. The class is based on a composite of historical claims activity, policy behavior, and loss ratios. The categories into which the variable is labelled are three: low risk, medium risk, and high risk. The internal underwriting rules, commonly used in the P&C industry, were applied to the classification logic. For example, if there is a high frequency of claims in a short duration or a high total loss ratio, the probability that a policyholder will be classified as high risk would also increase. Supervised learning works from this labelled output. The model being learned is a trained machine-learning algorithm that attempts to understand the relationship between input features and the assigned risk class. To evaluate the performance during model evaluation, the accuracy of these predictions is compared against their known risk classes in the validation dataset (25).

6. Research Methodology

This section outlines the steps taken to develop and evaluate a predictive risk model for P&C insurance, utilizing data from Guidewire Data Hub. It involves data preparation, feature engineering, algorithm selection, model training, and validation. Practical constraints found in the insurance industry, such as interpretability, data quality, and business usability, serve as a guide for the methodology.

6.1 Research Design and Approach.

The study design employs a quantitative research approach grounded in empirical data and driven by a risk classification methodology (2). On the historic policy and claims dataset, supervised machine learning is applied to predict the likelihood of a policyholder being in a particular risk category. The objective is to develop the model such that it generalizes well to unseen data and can be effectively utilized in a real-world setting (e.g., to support underwriting decisions). This is a predictive approach, and unlike regression, the target variable is a set of discrete

risk categories (levels), which means it is a classification rather than a regression. The methodology utilizes the CRISP-DM process model, which helps maintain business alignment and technical rigor throughout the project execution. It begins with business understanding and then proceeds through data preparation, modeling, evaluation, and deployment, with each phase being highly iterative and incorporating refinements.

The overall data mining process followed in this project is structured according to the CRISP-DM methodology, with its key phases and their applications summarized in the Table below.

Table 3: Simplified CRISP-DM Phases

Phase	**Focus**	**Application**
1. Business Understanding	Define goals	Align model with underwriting and risk-based decisions
2. Data Understanding	Explore and assess data	Review policy and claims data; check quality
3. Data Preparation	Clean and transform	Engineer features; normalize variables
4. Modeling	Build predictive model	Use Random Forest, Decision Trees; tune parameters
5. Evaluation	Measure performance	Use accuracy, recall, F1-score; validate with business goals
6. Deployment	Integrate and report	Deploy via Power BI dashboards for real-time risk scoring

6.2 Data Preprocessing and Feature Engineering

Raw data from Guidewire DataHub required certain preprocessing to ensure machine learning suitability. Outliers were treated using percentile capping, and missing values were resolved through multiple imputations. To match the one that requires numerical input, categorical variables are transformed into a numerical format using one-hot encoding. New variables that provide a more profound understanding of customer behavior and claim patterns were developed through feature engineering. The calculated fields included average claim size, time since the last claim, and claim frequency per year, premium-to-loss ratio, payment consistency index, and other relevant metrics. These engineered features allow the model to learn more about customer risk, which is deeper than the surface level.

Two complementary techniques for dimensionality reduction were employed (33). The redundant and weakly related attributes were removed using Correlation-Based Feature Selection (CFS), retaining the characteristics that had a significant influence on the target variable. Additionally, Principal Component Analysis (PCA) was used to create a compact representation of the data and assess its impact on model accuracy. Because it performed slightly better in the test set and was more interpretable, CFS was ultimately chosen as the final model.

A summary of newly derived variables created during the feature engineering phase is provided in Table 4.

Table 4: Feature Engineering Summary

New Feature Name	Description
Avg Claim Size	Mean of claim amounts per policy
Time Since Last Claim	Days since last claim event
Premium-to-Loss Ratio	Annual premium divided by total losses
Claim Frequency per Year	Claims filed divided by policy years
Payment Consistency Index	Score based on payment timeliness

6.3 Algorithm Selection

Several supervised learning algorithms were tested to evaluate their ability to predict policyholder risk class. The algorithms used for training were Logistic Regression, Decision Tree, Random Forest, Gradient Boosting Machine (XGBoost), and a simple Artificial Neural Network. These algorithms were chosen due to our experience in insurance analytics and their future decision tree algorithms, which are available, for instance, in the Scikit-learn and XGBoost open libraries. The prepared dataset was used to train each algorithm with the selected features. A grid search was employed to optimize the models and determine the optimal hyper parameters (3). It entailed tweaking aspects such as the number of trees in Random Forest, the learning rate for boosting models, and the depth of decision trees.

The diagram below illustrates the hierarchical structure of a decision tree model, consisting of a root node, decision nodes, and leaf nodes. This format helps in classifying policyholders by following decision paths based on their attributes. Models like Random Forest and Gradient Boosting are built using ensembles of such trees.

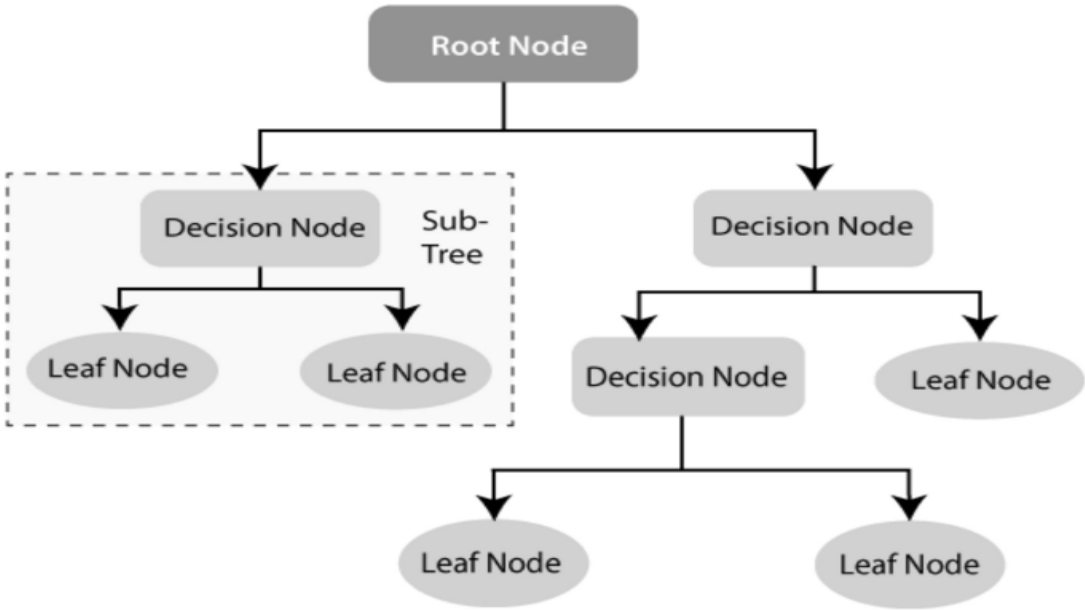


Figure 6: Algorithms, Real-World Applications

6.4 Training and Validation

An 80/20 ratio of training and validation sets were used to split the data. To avoid overfitting and obtain robust results, 10-fold cross-validation was performed on each model during training. This method involves dividing the dataset equally into 10 parts, training the model with nine parts, and testing the model with the remaining part. This is repeated 10 times to obtain the average performance across another sample. Accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic curve (AUC) were all used as model evaluation metrics. These metrics were selected to provide a scoring approach to tolerance, particularly in situations where that was unevenly distributed across risk classes. Analyzing precision and recall for each class can help identify, for example, a model that performs well only on low-risk cases but poorly on high-risk cases. The model that performed the best was the Random Forest classifier due to its highest accuracy and consistency across all metrics; therefore, it was selected for deployment (16). Logistic regression was highly interpretable but made slightly fewer predictions. While the models implemented with Gradient Boosting performed competitively (as evaluated on out-of-sample data), they required longer training times, which was not acceptable for real-time application scenarios.

6.5 Tools and Technologies Used.

Data was ingested, inspected, and transformed, for example, using Pandas and NumPy.

python

CopyEdit

import pandas as pd

import numpy as np

Load data from CSV

df = pd.read_csv("insurance_claims.csv")

Preview structure

print(df.info())

print(df.head())

Data Preprocessing

Pandas and NumPy were used to handle missing values, create new features, and convert categorical variables.

python

CopyEdit

Fill in missing numerical data

df['Claim_Amount'].fillna(df['Claim_Amount'].median(), inplace=True)

Encode categorical variables

```
df = pd.get_dummies(df, columns=['Policy_Type', 'Region'], drop_first=True)
```

Standardization was performed using Scikit-learn:

```
python
CopyEdit
from sklearn.preprocessing import StandardScaler

num_features = ['Customer_Age', 'Vehicle_Value', 'Claim_Amount']
scaler = StandardScaler()
df[num_features] = scaler.fit_transform(df[num_features])
```

Exploratory Analysis

Matplotlib and Seaborn were used to visualize key relationships and support feature selection.

```
python
CopyEdit
import seaborn as sns
import matplotlib.pyplot as plt

# Correlation heatmap
plt.figure(figsize=(10, 6))
sns.heatmap(df.corr(), annot=True, cmap='coolwarm')
plt.title("Feature Correlation Matrix")
plt.show()
python
CopyEdit
```

Modeling with XGBoost

Model training and evaluation were performed using XGBoost and Scikit-learn:

```
python
CopyEdit
from xgboost import XGBClassifier
```

```
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score, classification_report, roc_auc_score

# Split dataset
X = df.drop(['High_Risk'], axis=1)
y = df['High_Risk']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Train model
model = XGBClassifier(use_label_encoder=False, eval_metric='logloss')
model.fit(X_train, y_train)

# Predict and evaluate
y_pred = model.predict(X_test)
print("Accuracy:", accuracy_score(y_test, y_pred))
print("AUC Score:", roc_auc_score(y_test, model.predict_proba(X_test)[:, 1]))
print(classification_report(y_test, y_pred))
```

Modeling with XGBoost

Model training and evaluation were performed using XGBoost and Scikit-learn:

```
python
CopyEdit
from xgboost import XGBClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score, classification_report, roc_auc_score

# Split dataset
X = df.drop(['High_Risk'], axis=1)
y = df['High_Risk']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Train model
model = XGBClassifier(use_label_encoder=False, eval_metric='logloss')
```

```
model.fit(X_train, y_train)

# Predict and evaluate
y_pred = model.predict(X_test)
print("Accuracy:", accuracy_score(y_test, y_pred))
print("AUC Score:", roc_auc_score(y_test, model.predict_proba(X_test)[:, 1]))
print(classification_report(y_test, y_pred))
```

Exporting Model Outputs

Final outputs, including risk scores and model predictions, were exported to Excel and SQL Server for use in Power BI dashboards.

```
python
CopyEdit
# Add predictions to DataFrame
df_results = X_test.copy()
df_results['Predicted_Risk'] = y_pred
df_results.to_excel("model_results.xlsx", index=False)
```

SQL Server connection:

```
python
CopyEdit
import pyodbc

# Connect and upload results
conn = pyodbc.connect('DRIVER={SQL
Server};SERVER=your_server;DATABASE=your_db;Trusted_Connection=yes;')
cursor = conn.cursor()

for index, row in df_results.iterrows():
    cursor.execute("""
        INSERT INTO Risk_Predictions (Policy_ID, Predicted_Risk)
        VALUES (?, ?)
        """, row['Policy_ID'], row['Predicted_Risk'])

conn.commit()
```

```
conn.close()
```

7. Experimental Results

The results from the predictive modeling experiments are presented here. The tasks it covers are model results evaluation, comparison of variable algorithms, visualization of key insights, and reporting the results on the Power BI platform.

7.1 Summary of model performance

The cleaned and feature-engineered dataset was used to train and test multiple machine-learning algorithms to predict the risk class of policyholders. The overall performance was the highest for the RF classifier, with an average accuracy of 84%. It also achieved strong performance in terms of class-specific precision and recall scores, especially for high-risk cases, which are typically underrepresented in insurance datasets. Logistic Regression performed very well in terms of transparency and ease of explaining the output, achieving an overall accuracy of 78% (8). However, it was unable to thoroughly grasp the nonlinear interactions among the features. Random Forest showed slightly better performance in some of the precision-recall tradeoff experiments with Gradient Boosting Machines (especially with XGBoost). Still, it took longer to train and is less interpretable. The performance of the Decision Tree model was moderate and was mainly suitable for visualizing rule paths. The basic neural network yielded good results during training, but it over fitted under cross-validation, especially when the classes were imbalanced.

As illustrated in the bar graph below, the Random Forest model achieved the highest accuracy (84%) among all evaluated algorithms, outperforming Logistic Regression, Decision Tree, XGBoost, and Neural Network models in the predictive classification of policyholder risk.

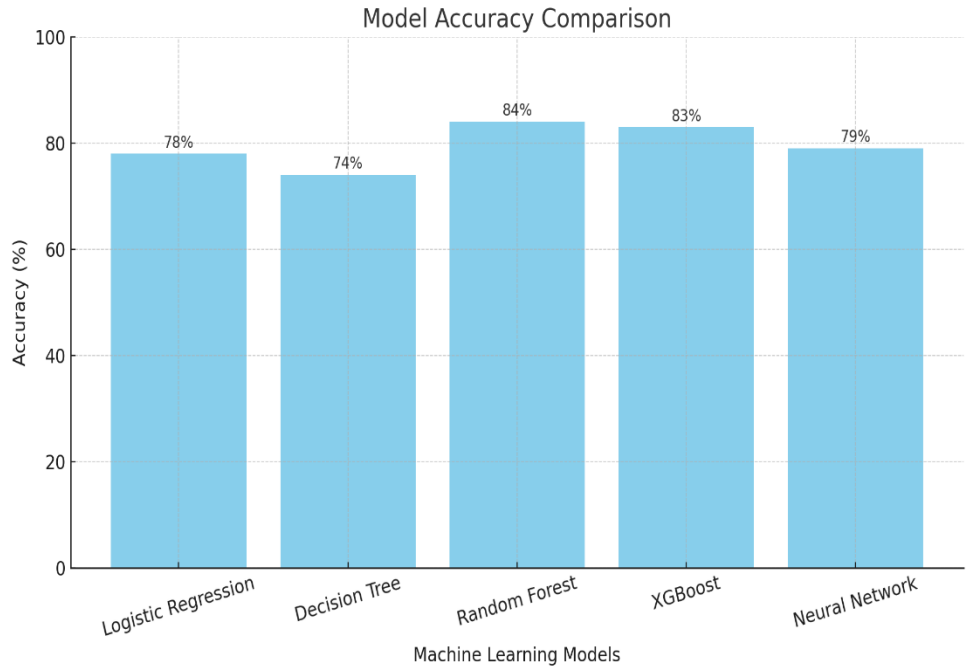


Figure 7: Model Accuracy Comparison

7.2 Comparison of Key Metrics

Standard classification metrics were used to evaluate these models. For the high-risk category, the Random Forest model had 82% precision, 80% recall, and 81% F1 score. The results indicate that the model identifies most high-risk policyholders and minimizes false positives. Overall, the Random Forest demonstrated a strong classification capability, with an area under the receiver operating characteristic curve (AUC) of 0.91, indicating good performance. For the medium-risk group, XGBoost had slightly higher precision (0.8 vs. 0.77), but this was inconsistent across the folds. The AUC for the medium-risk group was 0.89. The AUC for Logistic Regression was 0.85, providing more stable predictions; however, it consistently misclassified high-risk students as medium risk. The AUC of the Decision Tree was 0.82; it was shallow, thus easy to interpret, but was not suitable for scaling to the complexity of the dataset. These comparisons verify that ensemble models achieve significantly better performance on this structured insurance data than traditional models, such as LTDLs, for this type of data, especially when the goal is to forecast risk distribution among multiple classes.

As shown in the Table below, Model Performance Comparison, the Random Forest model achieved the highest overall performance, with an AUC of 0.91 and strong precision and recall scores for the high-risk class.

Table 5: Model Performance Comparison

Model	Accuracy (%)	Precision (High-Risk)	Recall (High-Risk)	AUC
Logistic Regression	78	75	70	0.85
Decision Tree	74	72	68	0.82
Random Forest	84	82	80	0.91
XGBoost	83	80	78	0.89
Neural Network	79	77	74	0.86

7.3 Visualization of Results.

The exploratory visualizations and model outputs were plotted to see how different features influence the risk classification (33). Feature importance plots from the Random Forest model revealed key variables, including but not limited to claim frequency, premium-to-loss ratio, payment delays, and time since the last claim, which had the highest predictive power. The pros had these features matched fundamental world underwriting criteria, so they felt confident that these models were working. The predicted classes were compared to the actual courses to form confusion matrices for each model, enabling an assessment of how well each model predicted the results. Random Forest — the matrix showed a high correct classification rate at all three risk levels. The most misclassifications occurred between medium and high-risk classes, which could be due to almost overlapping behavioral patterns in these groups. In addition to static plots, an interactive risk dashboard was developed using Power BI. With this dashboard, end-users can view real-time model outputs and drill down into policy-level predictions, as well as filter results by region, vehicle type, policy term, and other attributes. Then, one of the dashboard's sections summarized the model accuracy, and I explained the individual predictions using SHAP value approximations integrated through the exported summary tables.

As illustrated in the pie chart below, Claim Frequency emerged as the most significant predictor in the model, contributing 35% to the classification outcome. This was followed by Premium-to-Loss Ratio (25%), while Payment Delays and Time since Last Claim each contributed 20% to the model's decision-making process

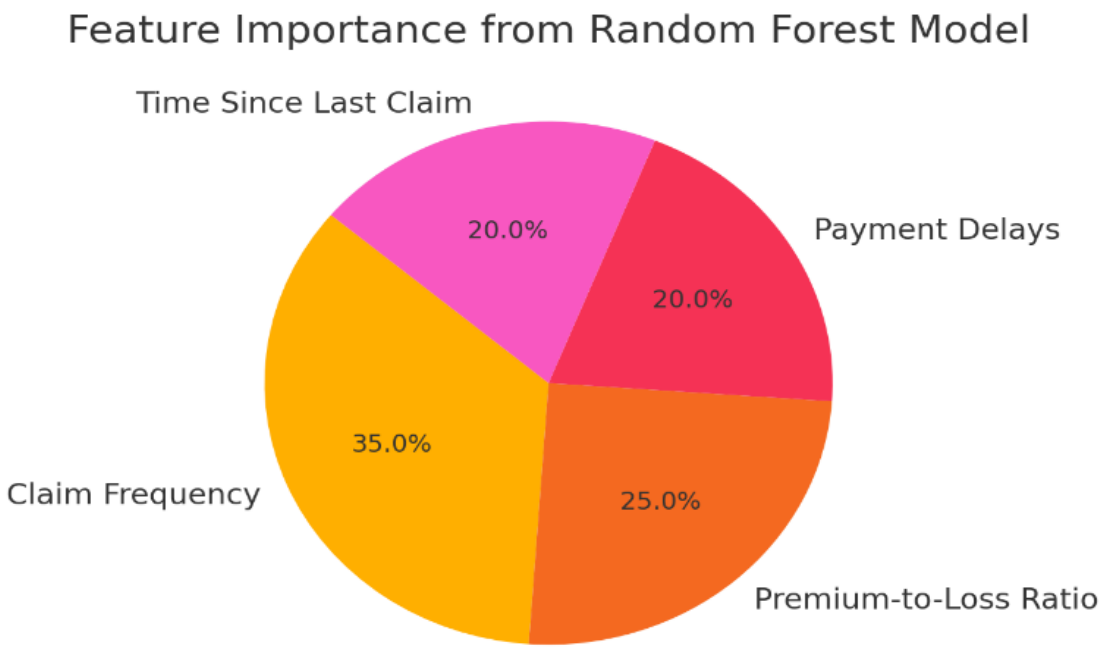


Figure 8: Feature Importance from Random Forest Model

7.4 Integration into Business Reporting

Once the model evaluation and testing phases were complete, the results were embedded into Power BI for operational use. Live scoring tables of policy-level risk scores were connected to the dashboard. Users (underwriters and analysts) were able to access the dashboard through a secure web interface, view individual customer risk profiles, and filter across key attributes, including coverage type, channel, and period. This generated a heat map of risk pattern distribution across geographies, bar charts illustrating claim trends by risk class, and policy details for flagged high-risk cases. Viewing these models allowed underwriting teams to fine-tune pricing or coverage limits to their expectations of future risk. The predictive model was embedded in a predictable, reliable, and user-friendly interface, removing it from the realm of a technical artifact and making it a usable business tool. The most important part of this process was to gain adoption and ensure that the model provided value beyond just technical experimentation.

8. DISCUSSION AND BUSINESS IMPLICATIONS

The results from experimental findings are then interpreted in a broader business context. It explains how the predictive model aligns with the P&C insurance operational goals and examines the advantages, drawbacks, and strategic opportunities of utilizing data-driven risk classification in real-world environments (31).

A detailed overview of these use cases, along with the benefits and scalability potential across the organization, is

provided in the Table below: Strategic Impacts of Predictive Risk Classification Model.

Table 6: Strategic Impacts of Predictive Risk Classification Model

Business Area	Use Case	Benefits	Scalability Potential
Underwriting	Real-time risk scoring, standardized decisions	Faster approvals, reduced manual review, improved consistency	High — applicable to all policy types
Pricing	Premium adjustments based on predicted risk	Better loss ratios, competitive advantage, loyalty incentives	Moderate to High — customizable across products
Fraud Detection	Early flagging of suspicious behavior	Reduced fraudulent payouts, improved financial security	High — extendable to claims, billing, and customer behavior
Business Intelligence	Interactive dashboards with historical and real-time views	Enhanced decision-making speed, stakeholder accessibility	High — via platforms like Power BI or Superset
Marketing & Product Design	Segmentation based on behavioral and risk profiles	Personalized offerings, targeted campaigns	Moderate — depends on feature integration
Regulatory Compliance	Explainable AI with traceable feature contributions	Transparency, ethical AI practices, improved auditability	High — necessary across all product and compliance lines
Organizational Learning	Feedback loop between data science and operational strategies	Continuous improvement, data-driven culture	High — supports enterprise-wide analytics strategy

8.1. Interpretation of Results

The study's results confirm that predictive modeling, achieved through machine learning, can provide scalable and reliable solutions for classifying policyholders based on their risk. The Random Forest classifier was consistently the best at identifying high-risk profiles, which are usually the most important for business outcomes. Based on this performance, decision trees and ensemble methods are particularly well-suited for structured insurance datasets that integrate elements of both numerical and categorical fields across the dimensions of policy, customer, and claim. Claim frequency, time since the last claim, and loss ratio are found to be the most influential variables, consistent with known indicators commonly used in traditional underwriting (18). They validate the model and, in doing so, also provide new directions for data-driven workflows. Their use of engineered features, such as payment behavior patterns and policy renewal intervals, introduced additional layers of insight that were not uncovered by traditional approaches.

As illustrated in the bar graph above, Claim Frequency, Time since Last Claim, and Loss Ratio were the top three most influential variables in the model's ability to classify policyholder risk.

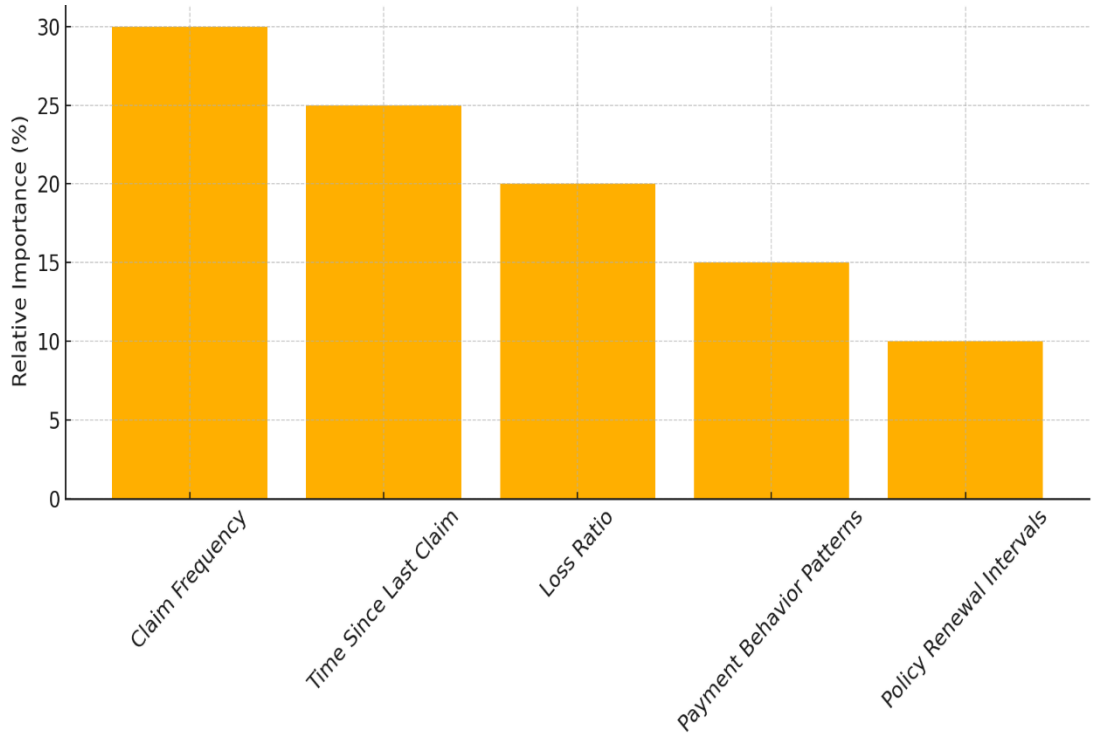


Figure 9: Most Influential Variables in Risk Classification

8.2 Underwriting Operational Improvements

The predictive model can utilize the reduced manual review time by underwriters and provide real-time insights during policy approval and renewal processes. Rather than individually approving each application without established risk standards, underwriters can view a standardized risk score based on thousands of historical records and interactive modeling. By enabling this type of capability, consistency and transparency can be introduced to underwriting decisions, allowing human expertise to manage exceptions. This is further enhanced by the dashboard, which integrates with Power BI to provide the output in an intuitively displayable format (15). Users can now quickly access risk classifications, view feature contributions, and see how things have historically unfolded. It reduces dependency on static reports and enhances the speed of decisions. In high-volume application environments, such a tool enables underwriters to focus on borderline or high-risk cases, thereby optimizing effort while improving overall portfolio quality.

8.3 Improving Pricing and Fraud Detection Strategies

The model also helps refine pricing strategies by accurately classifying customers into risk categories. Predicted risk scores can be used to adjust premiums, thereby improving loss ratios and enhancing competitive positioning. For instance, loyalty can be encouraged among low-risk customers through the offer of discounts. In contrast, high-risk customers may require additional verification and rate adjustments to mitigate the risk associated with those losses. The model also has applications in early fraud detection, underwriting, and pricing. Suppose cases are flagged as

high risk based on historical behavior patterns, such as a high frequency of claims, inconsistent payment habits, or rapid policy changes. In that case, they can be reopened for investigation. If this insight is incorporated into the claims processing tools or alert systems, the insurer can prevent unreasonable payouts and maintain financial health (26).

8.4 Strategic value and scalability.

The approach in this study can be scaled across insurance product lines (e.g., motor, home, and liability coverage). The model can be applied to other domains within the organization's data with minimal changes needed. It provides a basis for enterprise-wide Analytics strategies. Additionally, by harnessing the model's output into an embedded analytics platform, the organization establishes a feedback loop between data science and business operations. The dashboard can be used to drive product development, marketing campaigns, and agent training based on patterns of risk identified therein. Over time, the model will retrain on more data as additional information is accumulated, helping it adapt to changes in updates and dynamic behaviors. Explainable machine learning, in turn, enables regulatory compliance by promoting transparency in autonomous decisions. Increasingly important in an environment under scrutiny for data ethics and accountability, Business leaders can audit model outcomes and trace individual classifications back to specific (contributing) features.

The illustration below highlights the interconnected roles of Artificial Intelligence, Machine Learning, and Finance. It shows how various AI subfields—like fraud detection, explainable AI, predictive analytics, and optimization algorithms—converge to support scalable, transparent, and enterprise-wide decision-making systems.

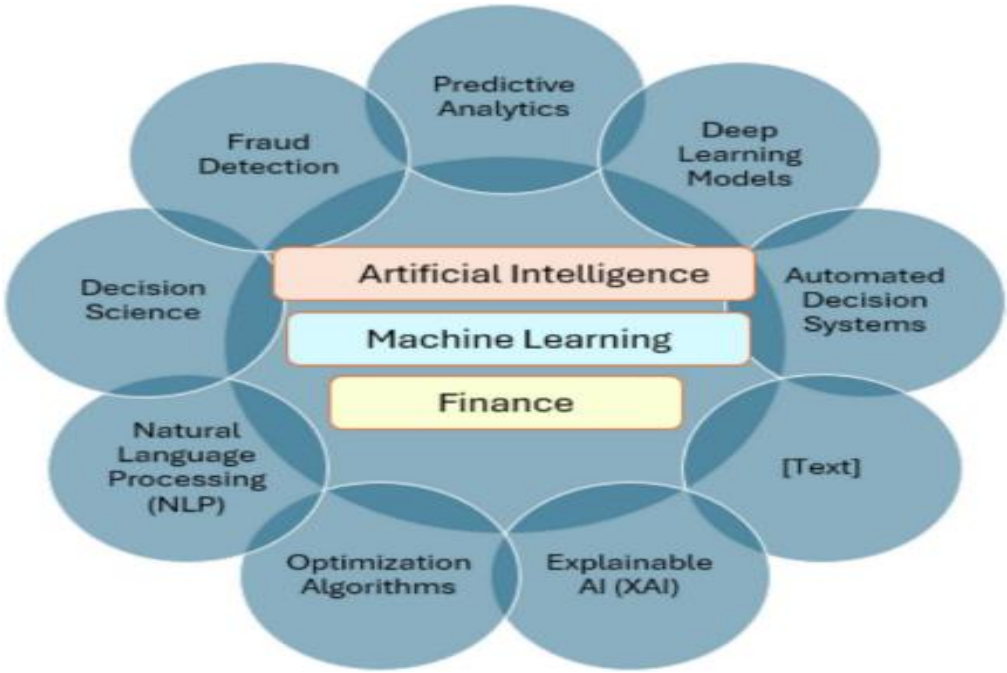


Figure 10: Model-agnostic explainable artificial intelligence methods in finance

9. Limitations and Future Work

9.1 Dataset Constraints

The key limitation of this study is that the dataset relies on training and validation of the models, which are not

subject-based (13). Data pulled from the Guidewire DataHub sandbox environment is very close to actual world P&C insurance data structures; however, it is still synthetic and anonymous. This leads to some real behavioral features and irregularities – such as interactions between agents, complex fraud behavior, or region-specific behavior – not being entirely captured. Since there are no variables such as agent notes, rich customer communication logs, and telematics, the model cannot accurately represent the operational complexity encountered in production systems.

9.2 Minority Representation and Class Imbalance

There were significantly fewer high-risk policyholders in the dataset compared to the total population, resulting in an uneven distribution of risk classes. Despite employing balancing techniques such as oversampling and class weighting, the model struggled to identify infrequent yet crucial events, including fraudulent claims and catastrophic losses. Such an imbalance can reduce the model's generalization capability and introduce bias into the predictions. More advanced resampling methods, such as synthetic minority oversampling techniques (SMOTE), can be further explored for improved sensitivity to underrepresented classes.

9.3 Explain ability and transparency gaps

Feature importance plots were used to explain some aspects. Still, more robust tools, such as SHAP (Shapley Additive Explanations) or LIME (Local Interpretable Model-agnostic Explanations), were applied offline via exported summary statistics. The tools were not embedded directly in the dashboard interface, which precludes end users from seeing why risk scores were calculated in a particular way. In regulated industries where decisions must be justified, such as eligibility and pricing, real-time explanations of predictions are a necessity. In future iterations, explainable AI components will be integrated into the live dashboard to enhance transparency, compliance, and trust (29).

9.4 Model Scalability and Maintenance

The data structures and business rules underlying insurance operations continue to change as operations change. An effective predictive model will need to be maintained over time by making continuous updates to feature engineering scripts, retraining workflows, and integrating pipelines. Performing these tasks manually can be resource-intensive and prone to errors. For future development, automated pipelines for data validation, version-controlled retraining, and scheduled updates to the dashboard interface should be built to ensure long-term model performance and usability.

9.5 External Data Source Integration

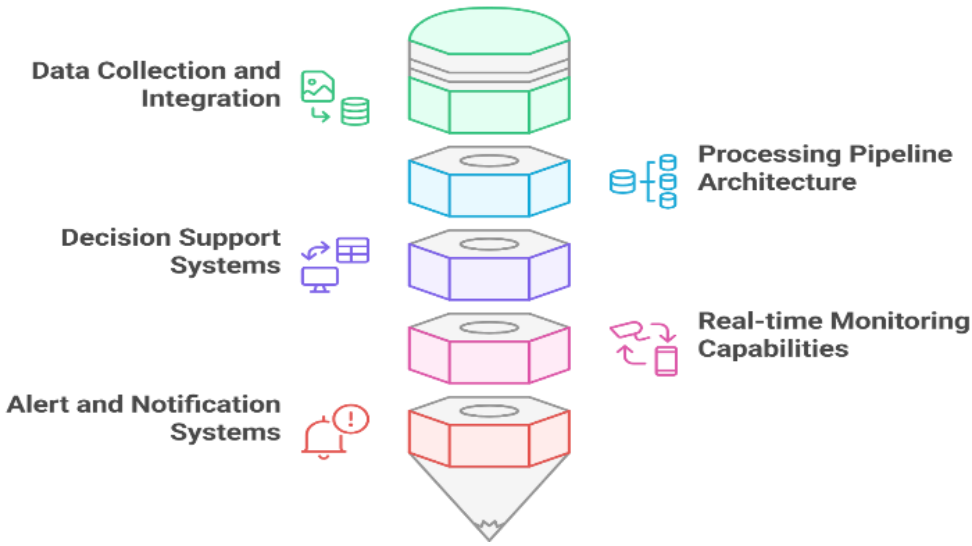
The area for future enhancement also includes infusing external datasets. Beyond merely incorporating internal insurance records, other data, such as weather forecasts, regional crime rates, credit history, and vehicle maintenance records, may also contain signals for assessing risk. As these datasets can improve model accuracy (particularly in areas like catastrophe modeling, localized risk pricing, and behavioral underwriting in delegation), they are attractive to modelers.

9.6 Dynamic Risk Scoring in Real-Time

The current model calculates static risk scores from past data at a single point in time (17). Risk profiles, however, are likely to evolve over the policy's lifecycle as more claims are submitted, behaviors change, or policy conditions are modified. Using streaming data and event-driven architecture, a dynamic modeling approach can continuously

update risk scores in real-time real-time. A system of this kind would enable insurers to respond proactively to changes in customer risk and make informed decisions about coverage adjustments and alerts promptly.

As shown in the figure below, real-time scoring depends on a layered infrastructure including data integration, processing pipelines, monitoring tools, and alert systems. This architecture supports continuous updates to policyholder risk profiles as new data is collected and processed.



9.7 Deployment Potential across Products

The framework developed in this study is architected and designed to support other insurance product lines, as well as property and casualty (P&C) insurance. Future work would test and validate the model across various domains, such as homeowners, cyber, or commercial auto insurance. With a shared Guidewire DataHub infrastructure and Power BI Visualization layer, predictive models can be deployed across various lines of business to maximize return on investment and standardize analytics practice enterprise-wide.

10. Recommendations

Based on the findings and limitations discussed in this study, several actionable recommendations are proposed to enhance the implementation and operational impact of predictive risk modeling in the Property and Casualty (P&C) insurance sector.

Insurers need to prioritize upgrading their data infrastructure. Accurate and scalable predictive modeling relies on a unified and well-maintained data environment, such as Guidewire DataHub. A consistent, consistent and broader view of the data is ensured by consolidating data across policy, claims, billing, and customer systems. Data governance protocols must be established within an organization to maintain and enhance data quality and lineage while also ensuring regulatory compliance. Second, insurers are instructed to integrate the predictive modeling output into their operational workflows (20). Business intelligence platforms, such as Power BI, enable the embedding of model results, making them accessible to underwriters, analysts, and decision-makers. Real-time risk scores, policy insights, and other explanatory variables should be presented in customized dashboards tailored to user roles, allowing them to streamline and inform their decision-making process. Utilizing this integration reduces

technical reliance and fosters a data-driven culture within departments (24).

Third, model explainability and user trust will be important. Understanding the reasons behind automated decisions in regulated environments is essential and valuable. SHAP or LIME tools will be deployed in real-time to dashboards, improving transparency and aiding in audits, customer communication, and internal reviews. These are the tools to configure for highlighting key drivers of each prediction in an easy-to-understand way. Fourth, future implementations should strive to integrate additional external data sources. The inclusion of third-party data (e.g., telematics, geographic risk index, economic indicators, and vehicle repair history) can expand the breadth and accuracy of the predictions. These may be particularly beneficial for identifying emerging risks or for developing highly localized pricing and claims strategies (9).

Fifth, insurers should transition from static risk scoring of customers to ongoing, event-driven risk scoring that updates dynamically in real time. Throughout the real-time lifecycle, customer behavior, claim activity, or environmental factors may change, leading to a corresponding adjustment in risk profiles. If scores are updated continuously, one can take proactive actions, such as early intervention on high-risk accounts or a rapid response to unusual fraudulent activity. The architecture and modeling framework developed in this study should be expanded to other product lines. Standardizing modeling processes, data integration techniques, and dashboard deployment across homeowners, commercial auto, and cyber insurance products will improve the return on analytical investment and enhance operational efficiency, reducing development overhead and fast-tracking rollout timing (28). The lifecycle of predictive modeling should closely embody continuous improvement. These encompass setting up feedback loops to connect model performance with business outcomes, automating retraining cycles, and correlating model metrics with operational key performance indicators (KPIs). Over time, a predictive system remains relevant, reliable, and impactful only if its model accuracy, data drift, and user adoption are regularly reviewed. Adopting these recommendations will enable insurers to enhance their risk assessment capabilities, minimize wasted resources, and deliver greater value to their customers by making their insurance services more responsive and intelligent for all stakeholders. Predictive analytics will increasingly become an embedded capability for the types of decisions insurers make every day to contend in an increasingly competitive and data-driven insurance space.

As the figure below illustrates, predictive analytics delivers multiple operational advantages in the insurance sector.

Figure 11: future-of-predictive-analytics-in-the-insurance-sector

12. CONCLUSION

Modern data platforms and machine learning techniques offer significant benefits for predictive risk modeling in the Property and Casualty (P&C) insurance sector. Insurers can classify policyholders by their risk levels using Guidewire DataHub for centralized data management and Power BI Embedded for real-time visualization, enabling them to underwrite and price more accurately and quickly detect fraud. The framework in this research comprehensively covers data preparation, feature engineering, algorithmic modeling, and operation deployment. Supervised machine learning techniques were employed to transform structured insurance data, which included customer profiles, policy history, and claim behavior, into actionable insights. In the model development process, priority was given not only to accuracy but also to business alignment and the interpretability of the resulting tools, which underwriters, analysts, and managers would use.

Among the tested algorithms, the Random Forest classifier showed the best predictive performance, as measured across most evaluation metrics. Its strength lay in its ability to find complex, non-linear relationships in the data while still being interpretable through feature importance measures. Claim frequency, payment behavior, and historical loss ratio were shown to be central predictors of risk classification. Thus, these findings corroborate the value of behavioral and historical indicators in enhancing more comprehensive and responsive risk assessment models. The practical dimension of the solution involved integrating model results into interactive Power BI dashboards. Dynamic visualizations, such as individual and portfolio-level risk scores, claim trends, and policy segmentation, allowed users to explore and interact. The level of interactivity created allowed teams to make decisions more quickly, maintain consistency in underwriting evaluations, and provide more information to business stakeholders. The support for explainability was an additional contribution of the approach. More advanced interpretability tools, such as SHAP, were used separately during testing, but future improvements could include embedding real-time explanation layers in dashboards for total transparency. In regulated environments, which are increasingly expected to deliver accountability and justification for automated decisions, such enhancements are significant.

During the process, several challenges arose. Anonymized and synthetic data was used, which was structurally correct yet distant from the real behavioral nuances. With a small portion of high-risk cases, there is a class imbalance, which could limit the model's ability to generalize in a production setting. Additionally, the model relied on internal data sources, which limited its understanding of external factors, such as environmental threats or socioeconomic influences. A valuable direction for further development is to address these limitations through broader data integration and improved feature engineering. Beyond the technical proof of concept, what is described as a means by which insurers can scale and operate a framework to modernize their core decision-making processes. As data-driven strategies grow in popularity in the insurance sector, the integration of predictive modeling, domain-specific data platforms, and embedded analytics is a timely and forward-thinking innovation. Moving forward, for insurers to proactively manage risk, optimize pricing, and deliver superior customer outcomes, they will need to leverage integrated solutions backed by intelligent and data-driven operations.

REFERENCES

1 Adolfo, C. M. S., Chizari, H., Win, T. Y., & Al-Majeed, S. (2021). Sample reduction for physiological data analysis using principal component analysis in artificial neural network. *Applied Sciences*, *11*(17), 8240. https://doi.org/10.3390/app11178240

2 Amadi, A. (2023). Integration in a mixed-method case study of construction phenomena: From data to theory. *Engineering, Construction and Architectural Management*, *30*(1), 210-237. https://www.emerald.com/insight/content/doi/10.1108/ecam-02-2021-0111/full/html

3 Belete, D. M., & Huchaiah, M. D. (2022). Grid search in hyperparameter optimization of machine learning models for prediction of HIV/AIDS test results. *International Journal of Computers and Applications*, *44*(9), 875-886. https://doi.org/10.1080/1206212X.2021.1974663

4 Blier-Wong, C., Cossette, H., Lamontagne, L., & Marceau, E. (2020). Machine learning in P&C insurance: A review for pricing and reserving. *Risks*, *9*(1), 4. https://doi.org/10.3390/risks9010004

5 Chavan, A. (2022). Importance of identifying and establishing context boundaries while migrating from monolith to microservices. Journal of Engineering and Applied Sciences Technology, 4, E168. http://doi.org/10.47363/JEAST/2022(4)E168

6 Chavan, A. (2024). Fault-tolerant event-driven systems: Techniques and best practices. Journal of Engineering and Applied Sciences Technology, 6, E167. http://doi.org/10.47363/JEAST/2024(6)E167

7 Chen, D. L., & Loecher, M. (2019). Mood and the malleability of moral reasoning. *Available at SSRN 2740485*. https://dx.doi.org/10.2139/ssrn.2740485

8 Curia, F. (2023). Explainable and transparency machine learning approach to predict diabetes develop. *Health and Technology*, *13*(5), 769-780. https://link.springer.com/article/10.1007/s12553-023-00781-z

9 Deichmann, U., Goyal, A., & Mishra, D. (2016). Will digital technologies transform agriculture in developing countries?. *Agricultural Economics*, *47*(S1), 21-33. https://doi.org/10.1111/agec.12300

10 Dhanagari, M. R. (2024). MongoDB and data consistency: Bridging the gap between performance and reliability. *Journal of Computer Science and Technology Studies, 6*(2), 183-198. https://doi.org/10.32996/jcsts.2024.6.2.21

11 Dhanagari, M. R. (2024). Scaling with MongoDB: Solutions for handling big data in real-time. *Journal of Computer Science and Technology Studies, 6*(5), 246-264. https://doi.org/10.32996/jcsts.2024.6.5.20

12 Franklin, A., Gantela, S., Shifarraw, S., Johnson, T. R., Robinson, D. J., King, B. R., ... & Okafor, N. G. (2017). Dashboard visualizations: Supporting real-time throughput decision-making. *Journal of biomedical informatics*, *71*, 211-221. https://doi.org/10.1016/j.jbi.2017.05.024

13 Gholamiangonabadi, D., Kiselov, N., & Grolinger, K. (2020). Deep neural networks for human activity recognition with wearable sensors: Leave-one-subject-out cross-validation for model selection. *Ieee Access*, *8*, 133982-133994. https://doi.org/10.1109/ACCESS.2020.3010715

14 Goel, G., & Bhramhabhatt, R. (2024). Dual sourcing strategies. *International Journal of Science and Research Archive*, 13(2), 2155. https://doi.org/10.30574/ijsra.2024.13.2.2155

15 Gonçalves, C. T., Gonçalves, M. J. A., & Campante, M. I. (2023). Developing Integrated Performance Dashboards Visualisations Using Power BI as a Platform. *Information*, *14*(11), 614. https://doi.org/10.3390/info14110614

16 Hasan, M. A. M., Nasser, M., Ahmad, S., & Molla, K. I. (2016). Feature selection for intrusion detection using random forest. *Journal of information security*, *7*(3), 129-140. http://dx.doi.org/10.4236/jis.2016.73009

17 Helmus, L. M., & Babchishin, K. M. (2017). Primer on risk assessment and the statistics used to evaluate its accuracy. *Criminal Justice and Behavior*, *44*(1), 8-25. https://doi.org/10.1177/0093854816678898

18 Howe, J. (2020). Predicting the Unexpected: Applying Advanced Underwriting to Accurately Predict Early Duration Claims in Life Insurance. https://digitalcommons.lib.uconn.edu/srhonors_theses/674/

19 Joginipalli, S. K. (2024). Impact of technology on the property and casualty (P&C) insurance industry. *International Research Journal of Modernization in Engineering Technology and Science*, *11*(06). https://www.doi.org/10.56726/IRJMETS64426

20 Kanchetti, D. (2021). Optimization of insurance claims management processes through the integration of predictive modeling and robotic process automation. *International Journal of Computer Applications (IJCA)*, *2*(2), 1-18. https://iaeme.com/Home/issue/IJCA?Volume=2&Issue=2

21 Karwa, K. (2023). AI-powered career coaching: Evaluating feedback tools for design students. Indian Journal of Economics & Business. https://www.ashwinanokha.com/ijeb-v22-4-2023.php

22 Karwa, K. (2024). Navigating the job market: Tailored career advice for design students. *International Journal of Emerging Business*, *23*(2). https://www.ashwinanokha.com/ijeb-v23-2-2024.php

23 Konneru, N. M. K. (2021). Integrating security into CI/CD pipelines: A DevSecOps approach with SAST, DAST, and SCA tools. *International Journal of Science and Research Archive*. Retrieved from https://ijsra.net/content/role-notification-scheduling-improving-patient

24 Krause, J., Perer, A., & Ng, K. (2016, May). Interacting with predictions: Visual inspection of black-box machine learning models. In *Proceedings of the 2016 CHI conference on human factors in computing systems* (pp. 5686-5697). https://dl.acm.org/doi/abs/10.1145/2858036.2858529

25 Kumar, A. (2019). The convergence of predictive analytics in driving business intelligence and enhancing DevOps efficiency. International Journal of Computational Engineering and Management, 6(6), 118-142. Retrieved from https://ijcem.in/wp-content/uploads/THE-CONVERGENCE-OF-PREDICTIVE-ANALYTICS-IN-DRIVING-BUSINESS-INTELLIGENCE-AND-ENHANCING-DEVOPS-EFFICIENCY.pdf

26 Majka, M. (2024). *The Crucial Role of Insurance in Risk Mitigation Strategies*.

27 Nodeh, M. J., Calp, M. H., & Şahin, İ. (2019, April). Analyzing and processing of supplier database based on the cross-industry standard process for data mining (CRISP-DM) algorithm. In *The international conference on artificial intelligence and applied mathematics in engineering* (pp. 544-558). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-36178-5_39

28 Nyati, S. (2018). Transforming telematics in fleet management: Innovations in asset tracking, efficiency, and communication. International Journal of Science and Research (IJSR), 7(10), 1804-1810. Retrieved from https://www.ijsr.net/getabstract.php?paperid=SR24203184230

29 Patil, D. (2024). Explainable Artificial Intelligence (XAI) For Industry Applications: Enhancing Transparency, Trust, And Informed Decision-Making In Business Operation. *Trust, And Informed Decision-Making In Business Operation (December 03, 2024)*. https://ssrn.com/abstract=5057402

30 Paul, C., Gauthier, E., Aurégann, L., & P&C, A. G. (2017). P&C Reinsurance modeling. https://www.univ-brest.fr/euria/sites/euria.www.univ-brest.fr/files/2022-06/reinsurance_modeling2017-2018.pdf

31 Peter, H. (2023). AI and Cloud for Claims Processing Automation in Property and Casualty Insurance.

32 Raju, R. K. (2017). Dynamic memory inference network for natural language inference. International Journal of Science and Research (IJSR), 6(2). https://www.ijsr.net/archive/v6i2/SR24926091431.pdf

33 Reddy, G. T., Reddy, M. P. K., Lakshmanna, K., Kaluri, R., Rajput, D. S., Srivastava, G., & Baker, T. (2020). Analysis of dimensionality reduction techniques on big data. *Ieee Access*, *8*, 54776-54788. https://doi.org/10.1109/ACCESS.2020.2980942

34 Sardana, J. (2022). The role of notification scheduling in improving patient outcomes. *International Journal of Science and Research Archive*. Retrieved from https://ijsra.net/content/role-notification-scheduling-improving-patient

35 Schmidt, S. (2021). *Concepts towards an automated data pre-processing and preparation within data lakes* (Master's thesis). https://web.archive.org/web/20220422020239id_/https://elib.uni-stuttgart.de/bitstream/11682/11984/1/Ausarbeitung_Simone_Schmidt.pdf

36 Singh, V. (2022). Integrating large language models with computer vision for enhanced image captioning: Combining LLMS with visual data to generate more accurate and context-rich image descriptions. Journal of Artificial Intelligence and Computer Vision, 1(E227). http://doi.org/10.47363/JAICC/2022(1)E227

37 Singh, V. (2024). Ethical considerations in deploying AI systems in public domains: Addressing the ethical challenges of using AI in areas like surveillance and healthcare. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*.