Data-Driven Decision Support for Auto Insurance: A Predictive Model for High-Cost Claims Management

Team: AutoCorrect

Team members: Hantian Chen, Yiwei Jiang, Anh Dang Phuong

Submission Date: November 1st, 2024

Contents

EXECUTIVE SUMMARY

This report aims to present a predictive model for identifying auto insurance claims exceeding $1000. The model was designed to provide accurate predictions of costly claims to enhance risk management strategies and optimize resource allocation for the ASNA Hackathon's participating insurance company. By incorporating the most influential variables in the dataset, our model achieved an accuracy of around 93.383%, offering reliable insights for business decisions.

This report is organized into six sections: **Introduction**, which outlines the objectives and context; **Model Selection and Explanation**, detailing the chosen model and key features; **Model Reliability and Validation**, demonstrating the model's accuracy and robustness; **Business Impact and Actionable Insights**, highlighting strategic recommendations; **Conclusion**, summarizing findings and future considerations; and an **Appendix** containing additional visuals and the modeling script. A **References** section is included at the end to support methodologies and concepts discussed throughout the report.

I. INTRODUCTION

1. Problem Statement

Insurance companies face challenges in accurately predicting high-cost claims, which affects financial planning and resource allocation. By identifying claims likely to exceed $1000, insurance companies can make informed decisions to manage risk and pricing, reducing unexpected financial liabilities.

2. Business Objective

The main business objective for predicting high-cost claims enables insurance companies to implement proactive and appropriate strategies such as adjusting premiums for high-risk customers, targeting retention rates, and allocating business resources more effectively.

3. Modeling Goal

Our predictive model aims to accurately identify auto insurance claims exceeding $1000. By utilizing historical data, the model's goal is to provide actionable insights for businesses.

II. MODEL SELECTION AND EXPLANATION

1. Data Overview

The provided train data set contains 29 features related to customer demographics, policy details, and claims history for 7290 past customers. Some of the key features include State, Customer Lifetime Value, Coverage Type, Employment Status, Policy Type, and Vehicle Size. The provided data set also contains some indexed indicators such as Coverage Index, Education Index, etc., which will be helpful for injecting numerical values into our predictive model.

	CustomerID	State	Customer Lifetime Value	Response	Coverage	Coverage Index	Education	Education Index	Effective To Date	Employment Status	...	Policy Type	Policy Type Index	Policy	Policy Index	Renew Offer Type	Sales Channel	Sales Channel Index	Vehicle Size	Vehicle Size Index	Claim over 1k
0	QC35222	California	3622.69	No	Basic	0	Bachelor	2	1/1/2024	Employed	...	Corporate Auto	1	Corporate L2	4	3	Web	0	Medsize	1	0
1	AE98193	Washington	10610.21	No	Basic	0	High School or Below	0	1/1/2024	Unemployed	...	Personal Auto	0	Personal L1	0	1	Branch	1	Medsize	1	1
2	TM23514	Oregon	13868.02	No	Extended	1	College	1	1/1/2024	Employed	...	Personal Auto	0	Personal L3	2	3	Web	0	Medsize	1	0
3	QZ42725	Washington	3119.69	No	Basic	0	Bachelor	2	1/1/2024	Unemployed	...	Personal Auto	0	Personal L3	2	2	Agent	2	Medsize	1	0
4	SG81493	Arizona	5999.04	No	Premium	2	Bachelor	2	1/1/2024	Employed	...	Corporate Auto	1	Corporate L1	3	2	Web	0	Medsize	1	0

5 rows × 29 columns

Figure 1. A Snippet of the Dataset that Contains 29 Features for 7290 Customers

2. Model

a. Model Assumptions

- **Data Accuracy**: The historical claims data provided is accurate, complete, and representative of typical claim behavior.

- **Independence of Features**: The model presumes that each feature has a direct influence on the outcome (whether a claim exceeds $1000) and that interactions between features do not have unexpected effects beyond what was captured in the training data.

- **Static Economic Conditions**: The model does not account for major economic shifts, such as recessions or inflation spikes, that could significantly change customer claim behavior. It assumes that the economic environment remains stable over time.

b. Model Choice: XGBoost Classifier

To effectively predict auto insurance claims that exceed $1000, we chose XGBoost for its high accuracy and efficiency in handling complex datasets with numerous features.

XGBoost was selected over a standard decision tree due to its higher accuracy and efficiency. Unlike a single decision tree, XGBoost uses boosting, which builds a sequence of trees where each tree learns from the previous one's errors, making it more accurate. It also includes regularization to prevent overfitting, which helps it perform better on new data.

XGBoost is highly efficient with parallel processing and handling missing values effectively, requiring less data preprocessing. These features make it ideal for large, complex datasets, offering precise, robust predictions for high-cost claims.

c. How the Model Works

Our XGBoost model functions as a sophisticated decision-making system that evaluates a variety of factors. It employs a method called boosting, which involves creating multiple decision trees in sequence. Each tree learns from the errors of the previous one, improving accuracy with each step. This process makes XGBoost highly effective at identifying patterns associated with high-cost claims.

- **Training with Historical Data**: This model is trained using past data on customer profiles, policy details, and claim outcomes. XGBoost learns from this data by building decision trees that divide customers into groups based on characteristics that influence claim size, such as Customer Lifetime Value, Income, and Coverage Level.

- **Iterative Learning Process**: XGBoost builds these trees sequentially, with each tree trying to improve on the mistakes of the previous ones. This iterative approach allows the model to capture complex patterns and subtle relationships that single decision trees might miss.

- **Prediction Output**: When a new claim is made, the XGBoost model evaluates it based on these learned patterns and produces a probability that the claim will exceed $1000. If the probability is high, the insurance team can prepare for a potentially high-cost claim and make data-driven decisions, like adjusting premiums or allocating resources more strategically.

d. Model Limitations

- **Sensitivity to Data Imbalance**: Our data is imbalanced, with far more claims under $1000 than over (see Appendix A1 for a visual representation). While we addressed this imbalance by adjusting class weights, the model can still favor the majority class (small claims), potentially overlooking some high-value claims. This may limit its effectiveness in accurately identifying all large claims, especially when the data is highly skewed.

- **Complexity and Interpretability**: Although XGBoost is accurate, it is also complex. Unlike simpler models, the inner workings of XGBoost are less transparent, which can make it challenging to interpret individual predictions. For stakeholders interested in understanding specific decisions, this lack of transparency might limit the model's usability.

- **Dependence on Historical Patterns**: XGBoost relies heavily on historical data patterns. If customer behavior or claim trends change significantly (e.g., due to economic factors or policy changes), the model may not adapt quickly, potentially reducing prediction accuracy.

- **Computational Requirements**: XGBoost requires more computing resources and time for training, especially when using large datasets or performing hyperparameter tuning. This can be a limitation when real-time or frequent model updates are needed.

e. Potential Improvements

- **Enhanced Data Balancing**: Currently, the model encounters far more small claims (<1k) than large ones (>1k), which can make it harder to identify high-cost claims accurately. We could explore advanced sampling techniques or undersampling the majority class. These methods would help balance the dataset more effectively, allowing the model to focus on high-value claims without being overwhelmed by the majority of small claims.

• **Refining Features for Deeper Insights**: By creating additional features—such as combining coverage level and customer value—we can capture more nuanced patterns in claim behavior. These refined features would allow the model to make even more precise predictions about high-cost claims.

• **Increasing Model Transparency**: Using interpretability methods to explain how each feature contributes to a prediction. This would provide valuable insights and help business partners understand the factors behind each prediction, building trust and supporting informed decision-making.

3. Feature Selection

Our model uses a selection of features that are particularly relevant for predicting whether an insurance claim will exceed $1000. These features were chosen based on both their predictive power and their practical significance in understanding customer behavior and risk. Initially, we identified the top features based on importance using a **Random Forest Classifier**, as shown in the chart below:

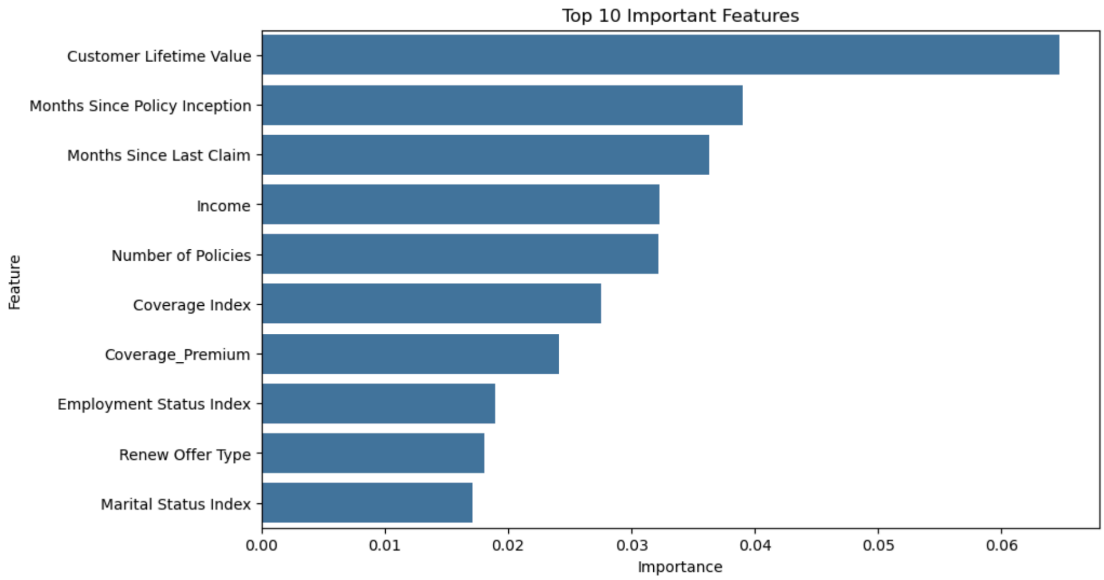


Figure 2. Top 10 Important Features

However, after closely analyzing the results, we noticed some overlaps and redundancies among the features that the model highlighted.

• **Overlap of Features**: While the model highlighted both **Coverage Index** and **Coverage Premium** as important features, these two variables are closely related. **Coverage Index** generally reflects the level of coverage (e.g., basic, standard, or premium), while **Coverage Premium** represents the actual cost paid for that coverage. These two features often capture similar risk information because higher levels of coverage correspond to higher premiums. To avoid redundancy, we selected **Coverage Index** as it encapsulates coverage scope directly without the added complexity of individual premium costs.

• **Irrelevance of "Months Since Policy Inception"**: The model initially identified **Months Since Policy Inception** as significant, but the length of time a customer has held a policy is not necessarily indicative of the likelihood or size of a claim. While tenure with the insurer may influence loyalty and retention, it does not strongly correlate with the likelihood or severity of claims. Actuarial analyses tend to focus more on recent behavior (e.g., recent claims or complaints) rather than policy duration alone. Therefore, we excluded this feature in favor of more behaviorally relevant indicators.

Based on this refined approach, we selected the following final features:

a. **Customer Lifetime Value (CLV)**

CLV measures the long-term value that a customer brings to the company, considering all policies held and the customer's loyalty over time. Higher CLV often correlates with customers who maintain multiple policies and exhibit loyalty. Actuarially, these customers are valuable but may have higher-value assets, increasing the likelihood of large claims.

b. **Months Since Last Claim**

This feature captures recent claims behavior, which is critical for assessing future claim risk. Customers with recent claims may be more likely to file additional claims, potentially including high-value claims. By comparison, customers with longer gaps since their last claim tend to exhibit lower risk. This feature is thus essential for estimating a customer's likelihood of future high-value claims.

c. **Income**

Income provides an indirect measure of a customer's financial standing and asset value.

Higher-income customers may own more valuable assets, such as high-end vehicles or properties, leading to potentially larger claims. Income also indirectly indicates a customer's ability to maintain and protect their assets. Compared to features like occupation or employment status, income is a more direct measure of asset value, making it particularly relevant for predicting high-cost claims.

d. **Number of Open Complaints**

This variable represents the number of unresolved complaints the customer has with the insurer.

Open complaints can signal dissatisfaction or unresolved issues, which may lead to higher claims activity. Customers with multiple complaints may feel entitled to claim more to offset perceived grievances. Compared to factors like customer engagement or policy type, open complaints directly reflect a strained customer relationship, which can correlate with an increased likelihood of large claims.

e. **Number of Policies**

This feature captures the total number of policies the customer holds with the insurer. Customers with multiple policies are generally more engaged with the insurer, reflecting a strong relationship. However, having more policies also means the customer has broader coverage, which could lead to higher exposure. Compared to other indicators of customer engagement, the number of policies gives a tangible measure of risk exposure, as it reflects the company’s overall liability with that customer.

f. Policy Type Index

This feature identifies the type of policy the customer holds, such as auto or home. Different types of policies carry distinct risk profiles; for example, auto policies may present higher claims frequency than life policies. Actuarial models use policy type as a primary segmentation factor to ensure accurate risk assessment and tailored pricing. Compared to coverage level, which applies within a policy type, policy type itself differentiates categories of risk, making it a fundamental feature for predicting claim size.

g. Coverage Index

The coverage Index represents the extent of coverage, from basic to extended levels. The coverage level is directly related to the claim size, as higher levels cover more extensive damage or losses. Customers with high coverage levels are statistically more likely to file larger claims. The coverage Index serves as a simplified measure of risk exposure, capturing both the customer’s willingness to pay for higher protection and the insurer’s potential liability.

h. Employment Status Index

Employment status is tied to financial stability, impacting a customer’s likelihood to file claims. Employed individuals typically have stable incomes, reducing the likelihood of large claims for financial compensation, while unemployed individuals might file larger claims to cover losses.

i. Renew Offer Type

The renewal offer type reflects the insurer’s customer risk assessment. Customers with standard or risk-adjusted offers may be considered higher risk, while loyalty offers suggest stability. It provides a valuable indicator of customer risk segmentation, helping to tailor pricing and predict future claims.

j. Marital Status Index

Marital status can affect risk behavior: married individuals, who may have dependents or shared financial responsibilities, tend to exhibit more cautious behavior and have a lower frequency of high-value claims. In contrast, single individuals may have different spending patterns or risk profiles that increase the likelihood of claims. By including marital status, we capture an important demographic factor that can affect claim behavior, allowing the model to better differentiate risk levels across customer profiles.

k. States Index (California Indicator)

This feature differentiates California residents from others due to California’s unique cost and regulatory factors.

Geographic differences significantly impact claim sizes due to variations in cost of living, local insurance regulations, and asset values. California, in particular, has a higher cost of living compared to many other states. As a result, customers in California may own higher-value assets, such as expensive vehicles or homes, which can lead to larger claim amounts.

III. MODEL RELIABILITY AND VALIDATION

1. Validation Metrics

To assess model performance, we used a range of validation metrics, including Accuracy, Precision, Recall, and F1-score. These metrics indicate the model's reliability, particularly its ability to correctly identify auto claims above $1000, making it suitable for business applications.

a. Accuracy

This model achieved an accuracy of roughly 93%, suggesting that the model correctly identifies whether a claim will exceed $1000 about 93% of the time. This high level of accuracy indicates strong overall performance.

b. Precision

• **For claims below $1000:** Precision was 0.94, meaning that when the model predicted a claim would be under $1000, it was correct 94% of the time. This high precision helps minimize the chance of mistakenly categorizing lower-cost claims as high-cost ones.

• **For claims above $1000:** For claims above $1000 (our target): Precision was 0.77, indicating that 77% of claims predicted to be above $1000 were indeed high-cost claims. While lower than the precision for claims below $1000, this is still valuable as it helps the business flag potentially costly claims with a reasonable degree of confidence.

c. Recall

• **For claims below $1000**: Recall was 0.98, meaning the model successfully identified 98% of actual lower-cost claims. This high recall for under-$1000 claims ensures we don't miss many claims in this category.

• **For claims above $1000:** Recall was 0.51, meaning the model correctly identified 51% of actual high-cost claims. While this score may appear low, it indicates the model's cautious approach. It avoids over-predicting high-cost claims, thus reducing the burden of investigating potential false positives.

d. F1-score

• **For claims below $1000**: The F1-score was 0.96, balancing the model's high precision and recall for these claims.

• **For claims above $1000:** The F1-score was 0.61, reflecting a moderate balance between the model's precision and recall for high-cost claims.

```
Confusion Matrix:
[[1276   24]
 [  78   80]]

Accuracy: 0.9300411522633745

Classification Report:
              precision    recall  f1-score   support

     Not >1k       0.94      0.98      0.96      1300
         >1k       0.77      0.51      0.61       158

    accuracy                           0.93      1458
   macro avg       0.86      0.74      0.79      1458
weighted avg       0.92      0.93      0.92      1458
```

Figure 3. A Snippet of Validation Metrics after Comparing the Results from the Model with the Real Results in the Train Dataset

The imbalance in the dataset likely affected the model's performance on claims above $1000. When the majority of entries (over 6000 out of 7290) fall below this threshold, the model is exposed to significantly fewer examples of high-cost claims during training, making it more challenging to accurately identify and predict them.

2. Business Reliability

While the model's metrics for predicting claims above $1000 are lower than those for claims below this threshold, they remain meaningful for the business because:

a. <u>Risk-Averse Approach</u>

Since costly claims can impact the company’s finances significantly, the model takes a conservative approach to predicting them. Lower recall (51%) for these claims means it may miss some high-cost claims but avoids a high rate of false positives, which reduces unnecessary actions based on incorrect predictions.

b. <u>Business Relevance</u>

Precision and recall together provide a reasonable detection system for costly claims, helping identify high-risk cases with 77% precision. This allows the business to prioritize review and resource allocation more effectively, focusing on claims that are more likely to exceed $1000.

c. <u>Practical Reliability</u>

The moderate F1-score for claims above $1000 signifies a balanced approach in prediction. Although some claims above $1000 are missed, the model provides reliable, actionable insights without excessive false alarms.

IV. BUSINESS IMPACT AND ACTIONABLE INSIGHTS

Leveraging the model’s reliable predictions, the company can achieve significant business benefits in terms of cost efficiency, customer engagement, and financial planning:

a. <u>Premium Adjustment for High-Risk Customers</u>

By focusing on the identified high-cost claims, the company can adjust premium pricing or apply risk surcharges to high-risk customers. For instance, customers with high coverage levels, multiple policies, or higher incomes tend to file costlier claims. Adjusting premiums for these profiles can help the company offset potential future losses.

b. Targeted Customer Engagement and Retention

Develop personalized engagement strategies for customers with moderate to high Customer Lifetime Value (CLV) or recent claims history. Offer these customers loyalty incentives, such as discounts for safe driving or policy bundles, to encourage retention and risk-reducing behaviors. For a high-value customer with no recent claims but a comprehensive coverage policy, offering a small discount for installing telematics devices (like vehicle tracking for safe driving) can promote safer behavior, potentially lowering future claims. By focusing on retaining high-value customers and promoting safer practices, the company can reduce the likelihood of high-cost claims while building long-term loyalty and enhancing the customer experience.

c. Resource Allocation for Claims Management

Use early identification of high-risk claims to strategically allocate resources toward investigating and managing these cases. High-risk claims, flagged by the model, should be prioritized for a more thorough review to prepare for potential high payouts. For example, a flagged claim from a high-risk customer can trigger an automated workflow for faster claims processing, with additional resources assigned to investigate potential fraud or high-cost recovery. This approach optimizes resource allocation, reduces the risk of delayed or inaccurate claims assessments, and ensures the company is better prepared for significant financial exposures.

V. CONCLUSION

This predictive model represents a strategic advancement in the company's approach to auto insurance claims management, reliably distinguishing between high- and low-cost claims. By enhancing claims processing efficiency and supporting informed resource allocation, the model strengthens the company's ability to manage claims proactively and cost-effectively.

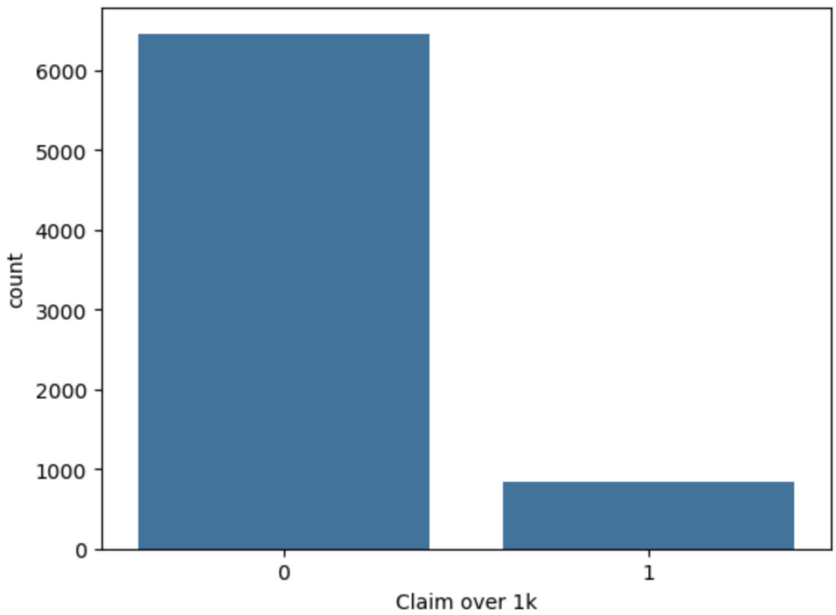
While the model excels in identifying lower-cost claims with precision, it also provides a foundation for continuous improvement in predicting high-cost claims. Future refinements in data handling and sensitivity to high-risk claims can further align the model with business objectives, enhancing the company's resilience and adaptability.

In summary, this model is a valuable, data-driven asset that empowers the company to make informed decisions, optimize resource use, and reinforce a more robust risk management framework, supporting the company's competitive advantage in a dynamic insurance landscape

VI. APPENDICES

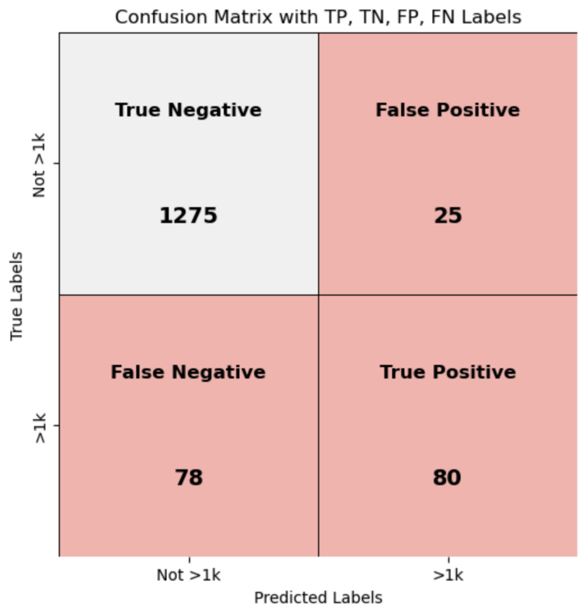
Appendix A: Visuals

A1. Distribution of Claims Above and Below $1000



This chart shows the distribution of claims classified as above or below $1000. The dataset has a significant imbalance, with the majority of claims falling below $1000. This imbalance impacts the model's training process, as it has fewer examples of high-cost claims to learn from, which can influence the model's recall for identifying high-cost claims. Addressing this imbalance is important for enhancing the model's performance in detecting high-value claims.

A2. Confusion Matrix for High- and Low-Cost Claim Prediction



This confusion matrix illustrates the model's performance in predicting claims above and below $1000. The matrix is divided into four sections: True Negatives (1275) and True Positives (80), which represent correct predictions, and False Positives (25) and False Negatives (78), which represent misclassifications.
The high number of True Negatives indicates the model's accuracy in identifying low-cost claims, while the False Negatives suggest opportunities for improving sensitivity toward high-cost claims. This balance between precision and recall informs strategic resource allocation and prioritization for high-risk cases.

Appendix B: Modeling Script

B1. Data Preparation and Cleaning

```
# Import necessary libraries
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import classification_report, confusion_matrix,
accuracy_score
import xgboost as xgb
# Load the dataset
df = pd.read_csv('train.csv')
# Create "States Index" where California is 1, other states are 0
df['States Index'] = df['State'].apply(lambda x: 1 if x == 'California'
else 0)
# Drop the original "State" column since we now have "States Index"
df = df.drop(columns=['State'])
# Check for missing values (optional: handle missing values if necessary)
print(df.isnull().sum())
```

B2. Feature Engineering and Encoding

```
# Convert categorical variables into dummy/indicator variables for other
columns
df_encoded = pd.get_dummies(df, drop_first=True)
# Define the features and target variable
X = df_encoded[['Customer Lifetime Value', 'Months Since Last
Claim','Income','Number of Open Complaints', 'Number of Policies', 'Policy
Type Index','Coverage Index','Employment Status Index', 'Renew Offer
Type','Marital Status Index', 'Education Index','States Index']]
y = df_encoded['Claim over 1k']
# Split into training and test sets
```

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=42)
```

B3. Model Training

```
# Train an XGBoost model
xgb_model = xgb.XGBClassifier(
    n_estimators=50,         # Number of boosting rounds
    max_depth=7,             # Depth of each tree
    learning_rate=0.1,       # Step size shrinkage
    random_state=42
)
xgb_model.fit(X_train, y_train)
```

B4. Evaluation and Metric Calculation

```
# Predict on the test set
y_pred = xgb_model.predict(X_test)
# Evaluate the model with various metrics
conf_matrix = confusion_matrix(y_test, y_pred)
accuracy = accuracy_score(y_test, y_pred)
class_report = classification_report(y_test, y_pred, target_names=["Not
>1k", ">1k"])
# Print evaluation metrics
print("Confusion Matrix:")
print(conf_matrix)
print("\nAccuracy:", accuracy)
print("\nClassification Report:")
print(class_report)
```

B5. Plotting Key Visuals

This section includes the code used to generate key visuals, such as the distribution of claims and the confusion matrix, which appear in Appendix A.

```
# Plot distribution of claims above and below $1000
sns.countplot(x='Claim over 1k', data=df)
plt.title("Distribution of Claims Over and Under $1000")
plt.xlabel("Claim over 1k")
plt.ylabel("Count")
plt.show()
# Plot confusion matrix with custom labels
import numpy as np
labels = np.array([["True Negative", "False Positive"], ["False Negative",
"True Positive"]])
colors = [["lightgreen", "yellow"], ["lightblue", "salmon"]]
```

```
plt.figure(figsize=(8, 6))
sns.heatmap(conf_matrix, annot=False, fmt='d', cbar=False,
xticklabels=["Not >1k", ">1k"], yticklabels=["Not >1k", ">1k"],
square=True, linewidths=0.5, linecolor='black')
for i in range(conf_matrix.shape[0]):
    for j in range(conf_matrix.shape[1]):
        plt.gca().add_patch(plt.Rectangle((j, i), 1, 1, fill=True,
color=colors[i][j], edgecolor='black', lw=0.5))
        plt.text(j + 0.5, i + 0.3, labels[i, j], ha='center', va='center',
color="black", fontsize=12, weight='bold')
        plt.text(j + 0.5, i + 0.7, f"{conf_matrix[i, j]}", ha='center',
va='center', color="black", fontsize=14, weight='bold')
plt.xlabel("Predicted Labels")
plt.ylabel("True Labels")
plt.title("Confusion Matrix with TP, TN, FP, FN Labels")
plt.show()
```

VII. REFERENCES

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering, 21*(9), 1263–1284. https://ieeexplore.ieee.org/document/5128907

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 4765–4774. https://doi.org/10.5555/3295222.3295230

Klugman, S. A., Panjer, H. H., & Willmot, G. E. (2012). *Loss models: From data to decisions* (4th ed.). Wiley.

Powers, D. M. W. (2011). Evaluation: From precision, recall, and F-measure to ROC, informedness, markedness, and correlation. *Journal of Machine Learning Technologies, 2*(1), 37–63.

Smith, V. K., & Desvousges, W. H. (1984). The generalized travel cost model and water quality benefits: A reconsideration. *Land Economics, 60*(4), 371–389. https://doi.org/10.2307/3146182