PREDICTIVE MODELING FOR POLICY LAPSE FORECASTING USING MACHINE LEARNING

Dissertation Presentation

- **Name**: Justin P Mathew
- **Supervisor**: Venkata Girish Kumar Nidra
- **Examiner**: Raghavendra G S

CONTENT

INTRODUCTION

An insurance policy lapse occurs when a policyholder fails to pay the premium within the grace period, leading to the termination of the policy.

This study focuses on predicting policy lapse within insurance datasets. By analyzing various factors such as payment modes, non-lapse guarantees, and policy status, we aim to build a robust predictive model.

IMPORTANCE OF PREDICTIVE MODELING

Predictive modeling is crucial in identifying potential lapses, enabling proactive interventions. Accurate forecasting helps in reducing financial losses, improving customer retention, and optimizing resource allocation.

Traditional management approaches are reactive, addressing lapses after they occur. This project aims to shift towards proactive management using predictive modeling.

OBJECTIVES

Objective 1

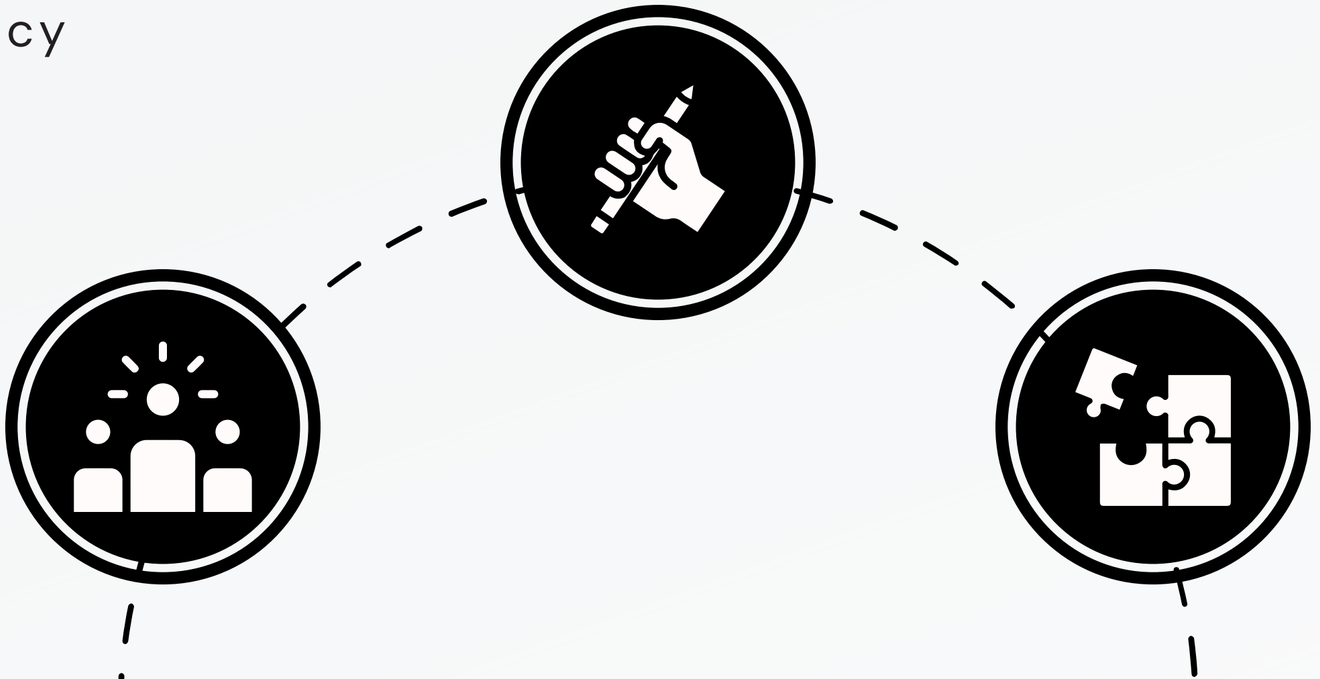
Develop a predictive model to accurately forecast policy lapses.

Objective 2

Identify key factors influencing policy lapses.

Objective 3

Evaluate performance of different predictive algorithms



DATA DESCRIPTION

- **Source of data**: Kaggle
- **Number of Rows and Columns**:
  - **Number of rows:** 185560
  - **Number of columns**: 22

- **Key features and attributes of the dataset**:
  - **CHANNEL1, CHANNEL2, CHANNEL3**: Distribution channels
  - **ENTRY AGE:** Age at which the policy was entered
  - **SEX**: Gender of the policyholder
  - **POLICY TYPE 1 - 3**: Types of policies

DATA DESCRIPTION (CONT)

- **Key features and attributes of the dataset**:
  - **POLICY STATUS:** Status of the policy (Inforce, Lapse, Expired, Surrender, Death)
  - **BENEFIT:** Benefit amount
  - **NON-LAPSE GUARANTEED**: Whether the policy has a non-lapse guarantee
  - **SUBSTANDARD RISK**: Substandard risk indicator
  - **NUMBER OF ADVANCE PREMIUM:** Number of advance premium payments made
  - **INITIAL BENEFIT:** Initial benefit amount
  - **Full Benefit?**: Whether the full benefit is applicable

DATA DESCRIPTION (CONT)

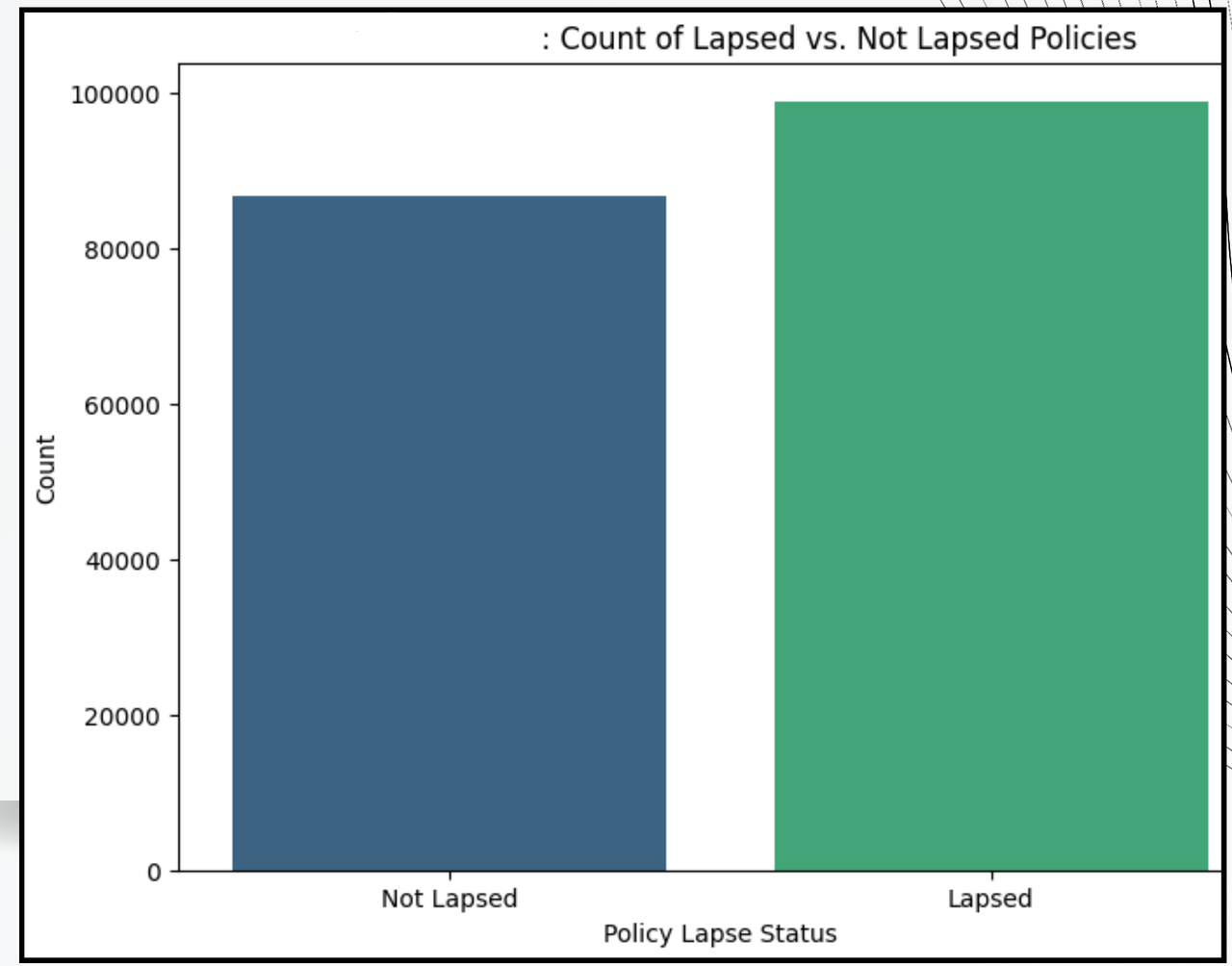
- **Key features and attributes of the dataset**:
  - **Policy Year (Decimal)**: Policy year in decimal
  - **Policy Year:** Policy year in integer
  - **Premium:** Premium amount
  - **Issue Date:** Date when the policy was issued
- **Data Types of Each Column:**
  - Mixed data types including integers, floats, and objects (strings).
- **Handling Missing Values:**
  - Columns Unnamed: 20 and Unnamed: 21 contain only null values and were dropped.

EXPLORATORY DATA ANALYSIS (EDA)

- **Summary Statistics:**
  - Mean, Standard Deviation, Min, Max, and Quartiles for numerical columns.
- **Box Plots:** Distribution of key numerical features such as ENTRY AGE, BENEFIT, Premium.
- **Correlation Heatmap**
  - **Purpose**: To identify the relationships between numerical features.
  - **Visualization**: Heatmap showing correlation values between features.
  - PAYMENTMODE_Monthly, NON LAPSE GUARANTEED_NLG Active, Policy Year, and PolicyYear(Decimal).

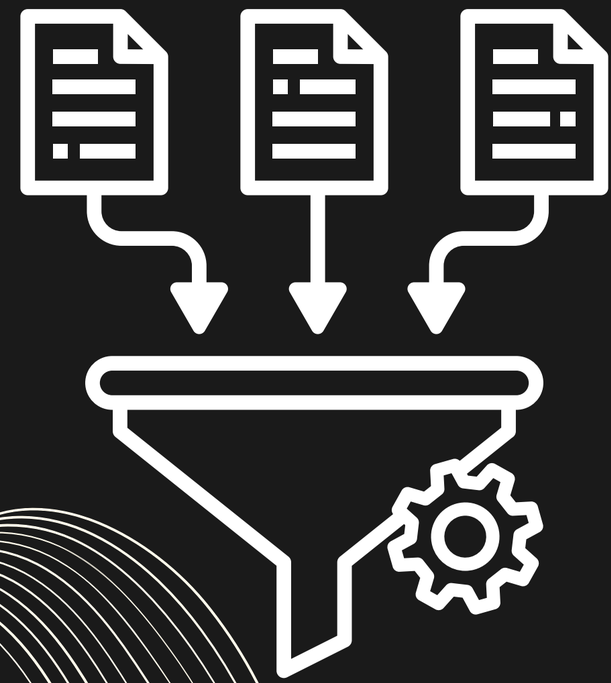
TARGET VARIABLE ANALYSIS

- **Creation of IS_LAPSE**: Binary target variable indicating policy lapse.
- **Class Distribution**: Bar chart showing the count of lapsed vs. not lapsed policies.



DATA PREPROCESSING

- **Encoding Categorical Variables**:
  - **OneHotEncoding**: Applied to PAYMENT MODE and NON LAPSE GUARANTEED.
  - **LabelEncoding**: Applied to SEX and Full Benefit.
- **Dropping Original Columns**: Removed POLICY STATUS after creating IS_LAPSE.



MACHINE LEARNING ALGORITHMS

- **Importance:**
  - **Predictive Power:** ML algorithms can identify patterns and make predictions with high accuracy.
  - **Automation:** Automates decision-making processes based on data.
  - **Scalability:** Can handle and process large volumes of data efficiently.
  - **Adaptability:** Models can be updated and improved as more data becomes available.

DECISION TREES

- A decision tree is a flowchart-like structure where each node represents a decision or a test, each branch represents the outcome, and each leaf node represents a class label (decision).

- **Advantages:**
  - **Easy to Understand and Interpret:** Simple to visualize and explain to stakeholders.
  - **Non-Linear Relationships:** Handles non-linear relationships well.
  - **Feature Importance:** Naturally ranks features by importance.

- **Disadvantages:**
  - **Overfitting:** Prone to overfitting, especially with complex trees.

RANDOM FORESTS

- A Random Forest is an ensemble of decision trees, usually trained with the "bagging" method.

- **Advantages:**
  - **Improved Accuracy:** Reduces overfitting by averaging multiple trees.
  - **Robustness:** Handles missing values and maintains accuracy.
  - **Feature Importance:** Provides insights into feature importance.

- **Disadvantages:**
  - **Complexity:** More complex than individual decision trees.
  - **Resource Intensive:** Requires more computational power and memory.

SUPPORT VECTOR MACHINES (SVMS)

- SVMs are supervised learning models used for classification and regression. They find the hyperplane that best separates the classes.

- **Advantages:**
  - **Effective in High-Dimensional Spaces:** Works well when the number of dimensions exceeds the number of samples.
  - **Versatile:** Can be used for both linear and non-linear classification (using kernels).

- **Disadvantages:**
  - **Training Time:** Can be slow to train, especially with large datasets.
  - **Complexity:** Less intuitive and harder to interpret than decision trees.

FUTURE WORK

Work to be done after Mid Sem before the final presentation

Performance Evaluation:

- Apply Decision Trees, Random Forests, and SVMs to the dataset.
- Evaluate and compare performance metrics (accuracy, precision, recall, F1 score).
- Identify which algorithm performs best in predicting policy lapses.

Refinement of Predictive Models

- **Hyperparameter Tuning**:
  - Further optimize model parameters for better accuracy and performance.
  - Techniques: Grid Search, Random Search, Bayesian Optimization.

THANK YOU