PREDICTIVE MODELING OF SURRENDER RATE IN LIFE INSURANCE

Prof. Nana Kena Frempong
(Supervisor)

Dagbey Thea	9348119
Amlalo Barnabas Doe	9344519
Ibrahim Hakim	9349719
Acheampong Kyei Rabbi	9342719

August 26, 2023

TABLE OF CONTENTS

BACKGROUND OF STUDY

- The life insurance industry provides policyholders with various options that can influence the insurer's exposure to liability or loss. Therefore, understanding the factors affecting policy surrender rates is essential.

- Machine learning is an artificial intelligence application that enables systems to automatically learn and improve from experience without the need for manual programming. (Arthur,1959)

PROBLEM STATEMENT

- A high surrender rate and how to determine at-risk policyholders is a significant problem faced by insurers. It affects both the policyholder and the insurer. Some of the effects include; loss of revenue, high expenses on the insurer, potential benefit loss, and reduced coverage on the part of the policyholder. To account for high uncertainty about policy surrender rate, ML techniques provide a predictive power to study the extent to which policyholders characteristics may influence the decision to surrender their policies. Hence the study utilizes three ML techniques to investigate this problem.

OBJECTIVES OF THE STUDY

- Explore and apply three ML techniques on the surrender data to estimate the probability of policyholder surrendering a life insurance policy
- Identify the key features that contribute to policy surrender.
- Identify at-risk policyholders who are likely to surrender their policies.

SIGNIFICANCE OF THE STUDY

- This study will provide more insight into the need for predictive modeling using machine learning algorithms and the key variables that contribute to policy surrender.

METHODOLOGY

- **Data collection**

  The Data used for the study was obtained from Kaggle, an online data platform. It was accessed on 22nd May, 2023 and the dataset consisted of 185,561 observations and 20 features.

- **Data pre-processing**

  Data pre-processing improves quality and refines datasets for model development by eliminating irrelevant features, converting relevant variables to numeric values, and then normalization was done.

METHODOLOGY CONT'D

- **Normalization**

  Data normalization involves arranging data uniformly, ensuring consistency across all records and fields. This means that the data should follow a standard normal distribution with a mean of zero and a variance of one N (0, 1). The formula is given below,

$$Z = \frac{x - \mu}{\sigma} \tag{1}$$

- **Feature Selection and Model Building**

  Feature selection involves identifying and choosing the relevant features that possess the highest predictive capability from a dataset while eliminating the irrelevant ones.The Models were built on 70% training dataset and evaluated on 30% testing dataset, with the Scikit-learn package.

METHODOLOGY CONT'D

- The Decision Tree Algorithm is a machine learning algorithm that constructs a tree-like model of decisions and their consequences by recursively splitting data into smaller subsets based on features
- The Random Forest method employs ensemble learning techniques for both classification and regression to generate a considerable number of decision trees.
- Support vector machines (SVMs) are supervised learning models and algorithms that effectively model complex problems like text classification, handwriting recognition, and complex number classification.SVMs are based on statistical learning frameworks.

- **Model Evaluation**

Table 1: Confusion Matrix for Surrender Prediction

Diagnosis	**In-force**	**Surrender**	**Lapse**
Predicted as In-force	TP	FP	FP
Predicted as Surrender	FN	TN	TN
Predicted as Lapse	FN	TN	TN

- **Algorithm Evaluation Metrics**

$$Accuracy = \frac{TN+TP}{TN+TP+FN+FP} \qquad F1 - measure = \frac{2 \times precision \times Recall}{Precision+Recall}$$

$$Recall/Sensitivity = \frac{TP}{TP+FN} \qquad Precision = \frac{TP}{TP+FP}$$

Data description and summary

POLICY STATUS	FREQUENCY	PERCENTAGES
Lapse	43253	52.60%
Surrender	7780	9.60%
Inforce	30894	37.60%
Death	230	0.30%
Expired	93	0.10%
TOTAL	**82250**	**100%**

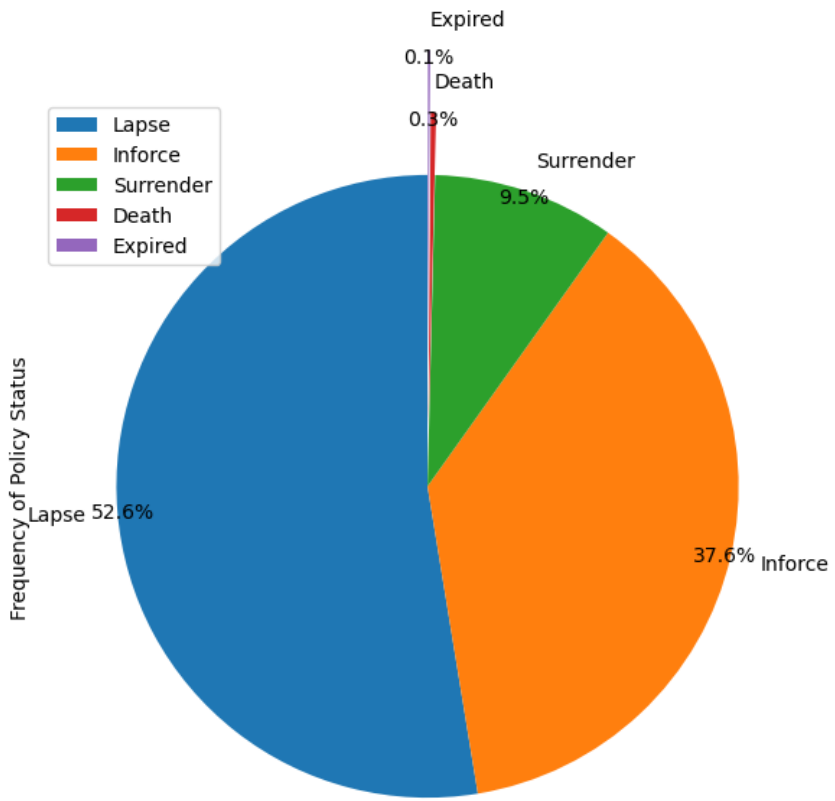


Figure 1: Frequency of policy status

Table 2: Data statistics before and after standardization

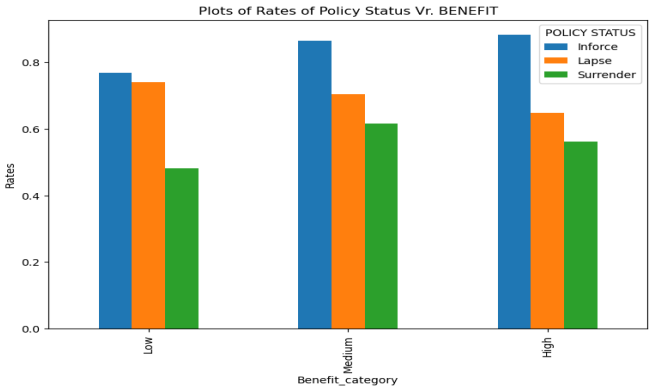
Before standardization

Features	Mean	Standard Deviation	Minimum	Median	Maximum
Policy Year (Decimal)	2.913	1.768	0	3	8
Policy Year	3.453	1.77	1	4	9
Premium	1289.998	5678.369	0	432	427053
Entry age	34.841	10.223	16	33	70

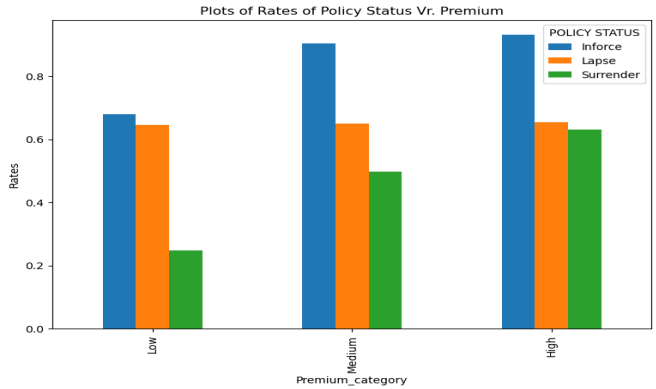
After standardization

Features	Mean	Standard Deviation	Minimum	Median	Maximum
Policy Year (Decimal)	0	1	-1.648	0.049	2.878
Policy Year	0	1	-1.386	0.309	3.134
Premium	0	1	-0.227	-0.151	74.98
Entry age	0	1	-1.843	-0.18	3.439

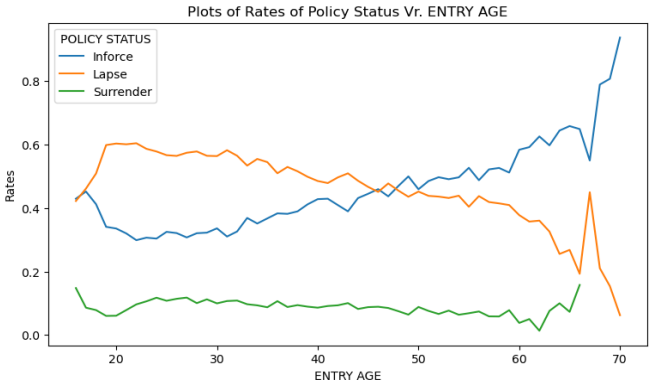
Bivariate Analysis.



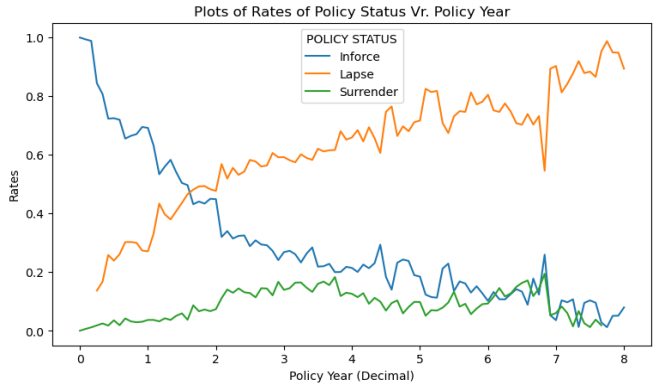
Benefit



Premium



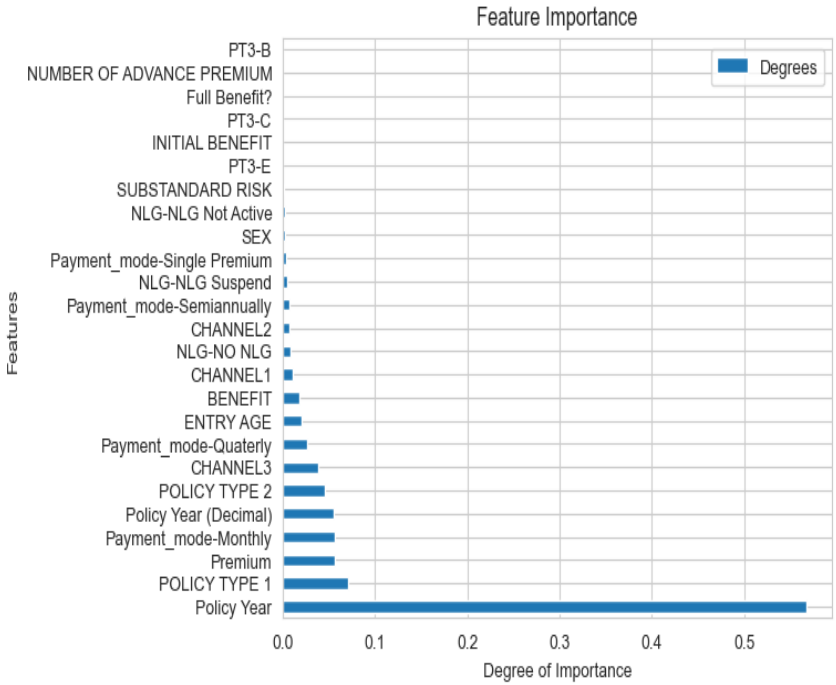
Entry age



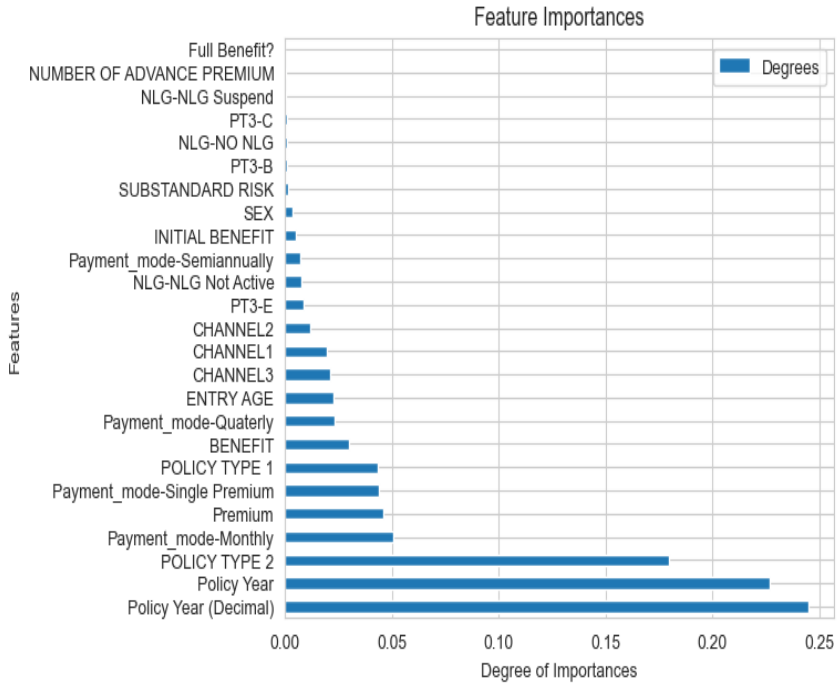
Policy year

Figure 2: Empirical hypothesis of features

Feature Importance



Decision Tree



Random Forest

Figure 3: Feature importance of Decision Tree and Random Forest

Model Evaluation

Table 4: Confusion Matrix for Machine Learning Algorithms

Decision Tree

Diagnosis	In-force	Surrender	Lapse
Predicted as In-force	5232	71	3966
Predicted as Surrender	335	116	1883
Predicted as Lapse	1647	27	11302

Random Forest

Diagnosis	In-force	Surrender	Lapse
Predicted as In-force	5236	37	3996
Predicted as Surrender	346	112	1876
Predicted as Lapse	1530	20	11426

Support Vector Machine

Diagnosis	In-force	Surrender	Lapse
Predicted as In-force	5327	13	3929
Predicted as Surrender	375	77	1882
Predicted as Lapse	1640	11	11325

Table 5: Algorithms performance evaluation

MODELS	ACCURACY	MSE	PRECISION	RECALL	F1 SCORE
DT	0.6774	0.3721	0.6729	0.6774	0.6442
RF	0.6825	0.3643	0.6893	0.6825	0.6481
SVM	0.6806	0.3667	0.6949	0.6806	0.6450

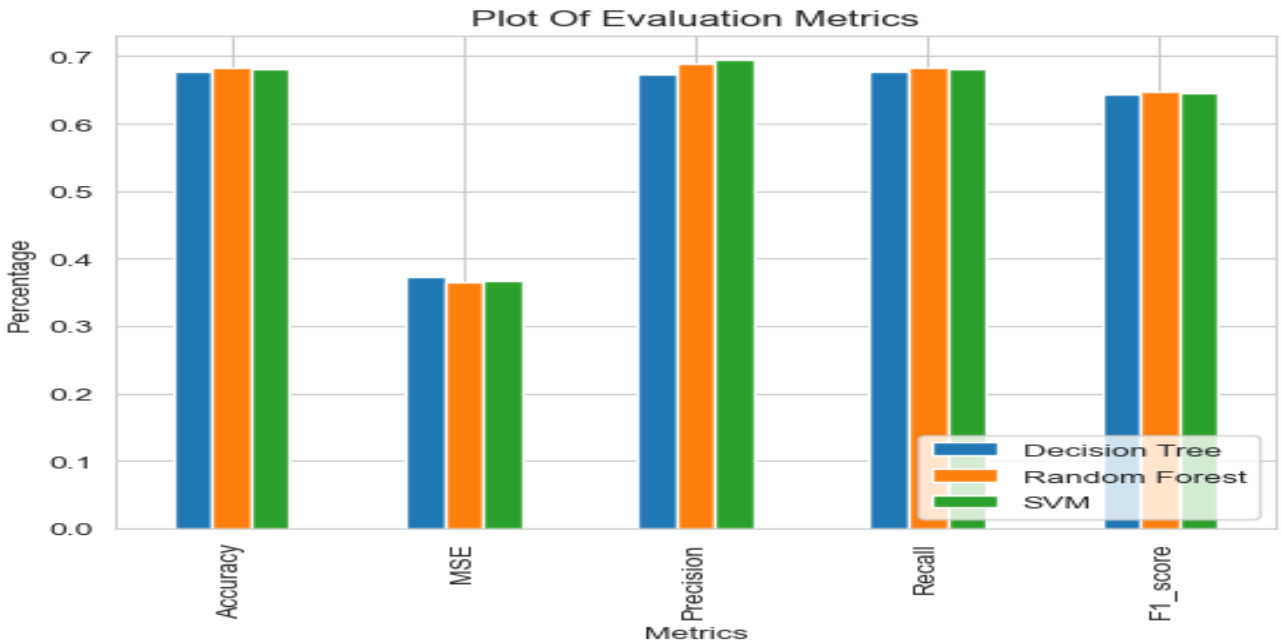


Figure 4: Graphical representation of evaluation metrics

CONCLUSIONS

- The research shows that Random Forest is the best model for accurate surrender predictions.
- Surrender rate is directly correlated with policy year and policy year (decimal) whiles policyholders who enter the scheme at age 30 and above have decreasing rate of surrender.
- Features of at-risk policyholders are higher premiums, an entry age of 30 and below, and longer policy years.
- The Study finds Random forest best for life insurance policy surrender status prediction; exploratory data analysis enhances model explainability.

RECOMMENDATIONS

- Life insurance companies should consider entry age, premiums and policy year before issuing a policy.
- Insurance companies and underwriters should apply machine learning techniques during their underwriting works to determine accurately at-risk policyholders .
- Machine learning experts should combine comprehensive data analysis and ML models to better understand the complex structure and dynamics within insurance data to improve surrender rate prediction.This can aid in tailoring retention strategies.
- Researchers, machine learning experts, and AI enthusiasts should consider comparing several machine learning techniques and other advanced methods in determining the most performing models before making their final decision.

THANK YOU.