introduction

Title: Predictive Analytics for Customer Purchases: Insights and Recommendations for Imperials Ltd

Objective: To apply data analytics techniques, particularly logistic regression and Random Forest models, to predict which consumers are likely to purchase life insurance products offered by Imperials Ltd.

Table of Contents

1. Introduction and Background

1.1 Introduction

The insurance industry is leveraging data analytics to enhance decision-making processes and tailor marketing strategies to better meet customer needs. This study aims to predict which consumers are likely to purchase life insurance by analyzing historical data from Imperials Ltd's customer database.

1.2 Background to the Problem

Predictive analytics in the insurance industry provides insights into customer behavior, risk assessment, and product optimization. This approach not only improves customer satisfaction by offering personalized products and services but also optimizes operational efficiency and profitability.

2. Literature Review

A review of literature related to life insurance prediction, machine learning algorithms, customer profitability forecasting, and the application of various data mining techniques in the insurance sector.

3. Methodology

3.1 CRISP-DM Process

The Cross Industry Standard Process for Data Mining (CRISP-DM) framework was used to address the business problem. The process includes business understanding, data understanding, data preparation, modeling, evaluation, and deployment.

3.2 Data Pre-Processing

The dataset comprises variables such as gender, education, house value, age, online purchase behavior, marital status, parenthood, occupation, mortgage bracket, homeownership, regional location, and family income grade.

3.3 Exploratory Data Analysis (EDA)

EDA was performed to identify trends and patterns. Key variables were visualized, and data quality issues were identified, such as missing values in education, undefined entries in gender, and outliers in house values.

3.4 Addressing Data Quality Issues

Data quality issues were addressed by:

- Imputing ‘U’ values in gender with ‘M’
- Replacing missing values in education with the mode
- Removing undefined entries in the child variable
- Imputing missing values in house_owner with the mode

3.5 Data Splitting Methodology

The dataset was split into 70% for training and 30% for testing, ensuring consistency through a predefined random seed.

3.6 Model Training and Selection

Two machine learning models were chosen:

- **Logistic Regression**: A statistical method for predicting binary outcomes.

- **Random Forest**: Constructs multiple decision trees to improve accuracy and reduce overfitting.

3.7 Model Evaluation and Comparison

Both models were evaluated and compared using metrics such as accuracy, precision, recall, F1 score, and AUC.

4. Results and Discussion

4.1 Logistic Regression Output

- Accuracy: ~67.57%
- Precision, Recall, and F1 Score: Similar values
- AUC: 0.74

4.2 Random Forest Model Output

- Accuracy: ~68.10%
- Precision, Recall, and F1 Score: Similar values
- AUC: 0.74

4.3 Comparison of Models

The Random Forest model slightly outperformed the logistic regression model in accuracy, precision, recall, and F1 score. Both models achieved an identical AUC value of 0.74, indicating comparable classification performance.

5. Conclusion and Recommendations

5.1 Conclusion

The analysis provided significant insights into predicting life insurance purchases. The Random Forest model demonstrated marginally better performance.

5.2 Recommendations

1. Adopt the Random Forest model for predictive tasks.
2. Enhance data quality for improved model performance.
3. Implement real-time analytics for dynamic marketing strategies.
4. Continuously gather and analyze customer feedback.
5. Foster a data-driven culture within the organization.

6. References

A comprehensive list of references used throughout the study.