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.

Selected Studies:

- Shamsuddin, S.N., Ismail, N., and Nur-Firyal, R. (2023). *Life Insurance Prediction and Its Sustainability Using Machine Learning Approach.*
- Others as listed in the document.

3. Methodology

3.1 CRISP-DM Overview

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 Business Understanding

The primary objective is to analyze customer reviews for insights that improve customer relationships and brand perception, helping strategic decisions to enhance customer satisfaction.

3.3 Data Understanding

The dataset includes reviews with brand name, textual content, star ratings, and emotional tags. Significant missing values in the 'Emotions' column necessitate special handling strategies.

3.4 Data Preparation

Data preparation tasks involve:

- Text Cleaning: Converting text to lowercase, removing punctuation, and URLs.
- Handling Missing Values: Employing imputation and semi-supervised learning methods to estimate missing emotional tags.
- Feature Extraction: Utilizing TF-IDF vectorization to convert text into a format suitable for machine learning analysis.

3.5 Modelling

Models selected include SVM, Logistic Regression, and Random Forests due to their proven effectiveness in text classification.

3.6 Evaluation

Models are evaluated using accuracy, precision, recall, and F1-score. Cross-validation ensures the models generalize well to unseen data.

3.7 Deployment

Deployment involves integrating the findings into a report detailing the analysis and offering actionable recommendations for the brands.

4. Data Analytics

Data analytics involves the use of various machine learning algorithms to analyze text data. Supervised and semi-supervised learning methods are employed to predict emotions from customer reviews.

Supervised Machine Learning

- Algorithms: SVM, Logistic Regression, Random Forest, Gradient Boosting, K-Nearest Neighbors, and Decision Trees.
- Vectorization: CountVectorizer and TfidfTransformer are used for text vectorization.

Semi-Supervised Machine Learning

- Self-Training Classifier is used to handle unlabelled data and improve model predictions.

5. Results and Discussion

5.1 Supervised Learning Analysis

Gradient Boosting was the best-performing model with 62.70% accuracy, 67.36% precision, and 62.95% F1 score. K-Nearest Neighbours performed poorly due to high-dimensional and sparse text data. SVM and Logistic Regression showed good performance, indicating their usefulness in text categorization tasks.

5.2 Analysis of Semi-Supervised Learning

Gradient Boosting also outperformed in semi-supervised learning with 61.11% accuracy and 60.71% F1-score. Random Forest performed better in this scenario, highlighting the benefits of using more unlabelled data to improve generalization.

5.3 Limitations of Supervised and Semi-Supervised Learning

- Supervised Learning: High-dimensional data, class imbalance, and overfitting challenges.
- Semi-Supervised Learning: Quality of unlabelled data, integration complexity, and computational resource intensity.

6. Conclusion and Recommendations

6.1 Conclusion

The analysis provided significant insights into predicting life insurance purchases. Gradient Boosting demonstrated high performance in both supervised and semi-supervised learning scenarios.

6.2 Recommendations

1. Adopt the Gradient Boosting 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.

7. References

A comprehensive list of references used throughout the study.