Insurance Purchase Prediction Using Machine Learning

By-Aditya Yadav

Date: 18th May 2024

Abstract:

This project focuses on predicting the likelihood of insurance purchase based on demographic factors, specifically age and income. The primary objective is to optimize customer targeting strategies for insurance companies. A dataset of 400 entries was curated and pre-processed, with features including age, estimated salary, and purchase status. Various machine learning algorithms were implemented and evaluated, with K-Nearest Neighbours (KNN) achieving the highest accuracy of 93%. The findings indicate that income is a critical predictor of purchase behaviour, and future work will explore the integration of additional variables to enhance model accuracy and robustness.

Table of Contents:

Introduction:

The ability to accurately predict whether a customer will purchase insurance is critical for insurance companies to optimize their marketing and sales strategies. This project aims to develop a machine learning model that can predict insurance purchase likelihood based on key demographic factors such as age and income. By leveraging various machine learning algorithms, the project seeks to identify the most effective method for predicting purchase behaviour, thereby enabling more targeted and efficient customer outreach.

Literature Review:

The prediction of insurance purchase behaviour has been a subject of interest in both academic and industry settings. Previous research has explored the use of demographic factors, such as age and income, as predictors of insurance purchase likelihood. Machine learning techniques, including logistic regression, decision trees, and ensemble methods, have been widely used in similar predictive modelling tasks. However, challenges remain in balancing model accuracy with interpretability and scalability in real-world applications.

Problem Statement:

The primary objective of this project is to predict whether a customer will purchase insurance based on their age and income. This problem is critical for optimizing customer targeting strategies. The project assumes that the dataset is representative of the broader population and that age and income are sufficient predictors of purchase behaviour. Limitations include the potential influence of other unaccounted demographic or behavioural variables.

Data Collection and Preprocessing:

The dataset used in this project comprises 400 entries, with features including age, estimated salary, and purchase status (0 for no purchase, 1 for purchase). Data cleaning involved handling missing values and standardizing features using StandardScaler to ensure consistency across different scales of data. The dataset was then partitioned into training (75%) and test sets (25%) to facilitate model training and evaluation.

Methodology:

The following machine learning algorithms were implemented using Python's Scikit-learn library:

- **Logistic Regression:** Achieved an accuracy of 89%, serving as a baseline model.
- **K-Nearest Neighbours (KNN):** Delivered the highest performance with 93% accuracy, leveraging proximity-based classification.
- **Support Vector Machine (SVM):** Achieved an accuracy of 90%, demonstrating robust classification capabilities.
- **Decision Tree:** Achieved 91% accuracy, providing insights into feature importance.
- **Random Forest:** Also achieved 91% accuracy, highlighting the benefits of ensemble learning.

The selection of these algorithms was based on their suitability for classification tasks and their ability to handle the given dataset.

Implementation:

The models were implemented using Python and the Scikit-learn library. The dataset was split into training and test sets, and each model was trained on the training set. The code implementation involved defining the models, fitting them to the training data, and evaluating their performance on the test set using accuracy scores and confusion matrices. The following GitHub link provides access to the full implementation: [GitHub link].

Results:

The results of the model evaluations are summarized below:

- **Logistic Regression:** 89% accuracy
- **K-Nearest Neighbours (KNN):** 93% accuracy
- **Support Vector Machine (SVM):** 90% accuracy
- **Decision Tree:** 91% accuracy
- **Random Forest:** 91% accuracy

Visualizations created using Matplotlib illustrate the distribution of predicted vs. actual outcomes, providing a clear comparison of model performance.

Discussion:

The results indicate that income is a significant predictor of insurance purchase behaviour across all models. KNN outperformed other algorithms, suggesting that proximity-based classification is particularly effective for this dataset. While the models generally performed well, there were some unexpected outcomes, such as the similarity in performance between Decision Tree and Random Forest models, which warrants further investigation. The strengths of the models include their high accuracy and interpretability, while limitations involve the potential for overfitting and the need for additional demographic or behavioural data.

Conclusion:

This project successfully developed and evaluated multiple machine learning models to predict insurance purchase likelihood based on age and income. The findings demonstrate the importance of income as a predictor and suggest that KNN is the most effective algorithm for this task. Future work will focus on incorporating additional variables to enhance model robustness and exploring deployment options for real-time applications in insurance marketing and sales.