Report on Optimizing Subsidy Delivery Using Feature Selection

1. Introduction

In the contemporary landscape of socio-economic development, the efficient allocation and delivery of subsidies are crucial for ensuring that financial assistance reaches the intended beneficiaries. Misallocation of subsidies can lead to significant economic inefficiencies and social inequities. In this context, the integration of machine learning models, particularly those utilizing feature selection, has emerged as a powerful tool in optimizing subsidy delivery mechanisms. This report delves into the development and implementation of an optimized subsidy delivery model, utilizing univariate and bivariate analysis to identify key features. By focusing on critical variables and employing classification algorithms such as Logistic Regression and K-Nearest Neighbors (KNN), the model achieves a high accuracy in predicting salary status—a proxy indicator for eligibility, thereby ensuring a more effective subsidy distribution.

2. Problem Statement

The core challenge in subsidy distribution lies in accurately identifying eligible beneficiaries. Traditional methods often suffer from inefficiencies due to the inclusion of irrelevant features, leading to misallocation of resources. Inaccurate predictions about salary status can result in subsidies being granted to ineligible individuals or, conversely, withholding support from those who genuinely need it. Therefore, the primary objective of this project is to develop a robust machine learning model that optimizes subsidy delivery by identifying and focusing on the most relevant features, thus improving prediction accuracy.

3. Methodology

The methodology for this project is divided into several key stages:

3.1. Data Exploration and Preprocessing:

The dataset used in this project is derived from income-related records, encompassing various demographic and economic features. The first step involved performing exploratory data analysis (EDA) to understand the distribution of variables, identify missing values, and assess correlations among numeric features.

- Handling Missing Data:

The dataset initially contained 1,816 rows with missing values. These were primarily in the 'Job Type' and 'Occupation' columns, which were crucial for determining income status. Missing values were removed to ensure the integrity of the dataset.

- Categorical and Numeric Analysis:

Descriptive statistics revealed that 75% of individuals reported zero capital gains, indicating a significant skew in the data. Categorical variables such as 'Job Type' and 'Occupation' were examined, revealing that the majority of people fell into the 'Private' job category and earned less than $50,000 annually. These insights informed the subsequent feature selection process.

3.2. Feature Selection:

Univariate and bivariate analyses were conducted to identify key features that influence salary status. The following features were considered critical based on their predictive power:

- Age:

A significant correlation was observed between age and salary status, with individuals aged between 25 and 45 showing higher frequency distributions.

- Education Level:

Higher education levels, particularly those with a bachelor's, master's, or doctoral degree, were strongly associated with higher salary brackets.

- Marital Status:

Married individuals, especially those in a 'Married-civ-spouse' relationship, showed a significant variance in salary distribution, making this a crucial feature.

- Occupation:

Occupations such as 'Exec-managerial' and 'Prof-specialty' exhibited an almost equal distribution between higher and lower salary brackets, indicating their importance in predicting income levels.

- Hours Worked Per Week:

Individuals working more than 50 hours per week were more likely to earn over $50,000, further justifying the inclusion of this feature.

Features such as 'Gender', 'Native Country', 'Race', and 'Capital Loss' were found to have minimal impact on salary prediction and were subsequently excluded from the model to avoid overfitting and reduce computational complexity.

3.3. Model Development:

Two primary classification algorithms were employed: Logistic Regression and K-Nearest Neighbors (KNN). These models were trained and tested using a 70-30 train-test split on the processed dataset.

- Logistic Regression:

Initially, a Logistic Regression model was implemented with all available features. After feature selection, the model was retrained using only the key features identified in the previous step. The accuracy of the Logistic Regression model with all features was approximately 83.6%. Despite reducing the number of features, the model maintained a similar accuracy level, demonstrating the effectiveness of the feature selection process.

- K-Nearest Neighbors (KNN):

KNN was tested with varying values of 'n' (the number of neighbors). The optimal performance was observed at `n=20`, achieving an accuracy of approximately 84.7% when all features were included. Notably, even with a reduced set of features, KNN achieved a comparable accuracy of approximately 83.6%, further validating the selected features' relevance.

3.4. Evaluation Metrics:

Model performance was evaluated using metrics such as accuracy score and confusion matrix. The confusion matrix provided insights into the number of correctly and incorrectly classified instances, allowing for the assessment of the model's precision in predicting salary status.

4. Results

The feature selection process was pivotal in enhancing the model's efficiency while maintaining high accuracy. The Logistic Regression model, post feature selection, achieved an accuracy of approximately 83.6%, demonstrating that the exclusion of less relevant features did not compromise predictive power. Similarly, the KNN model, with `n=20`, achieved a slightly higher accuracy of approximately 84.7% when all features were considered. However, even with a reduced feature set, the model's accuracy remained robust at approximately 83.6%. This marginal difference suggests that the selected features were sufficient for accurate prediction, reinforcing the validity of the feature selection process.

5. Conclusion

The project successfully demonstrated the importance of feature selection in optimizing subsidy delivery models. By focusing on critical variables such as age, education level, marital status, occupation, and hours worked per week, the model was able to achieve high accuracy in predicting salary status. The use of both Logistic Regression and K-Nearest Neighbors provided a comparative analysis, highlighting that a reduced feature set could maintain, or even enhance, model performance.

This study underscores the potential of machine learning in refining subsidy distribution mechanisms, ensuring that resources are allocated efficiently and equitably. Future work could explore the integration of additional features or the application of more complex models, such as ensemble methods, to further improve prediction accuracy and model robustness. Nonetheless, the current model serves as a robust foundation for optimizing subsidy delivery, with significant implications for socio-economic policy and resource management.

In conclusion, the integration of feature selection with machine learning models offers a powerful approach to addressing the challenges of subsidy allocation. By focusing on the most relevant variables, the model not only improves accuracy but also ensures that subsidies are delivered to the right individuals, minimizing the risk of misallocation and enhancing the overall effectiveness of financial assistance programs.