# Loan Prediction Machine Learning Model Report

## Introduction

Loan prediction is a pivotal aspect of the financial industry, providing crucial insights for financial institutions to assess customer behavior and make informed lending decisions. This project leverages data analytics and machine learning to predict loan eligibility based on consumer behavior, aiming to improve accuracy and reliability in loan predictions.

## Dataset Information

The dataset comprises various features reflecting user demographics, professional experience, ownership status, marital status, and risk flags. These features are used to predict whether a consumer is worthy of a loan.

## Dataset Columns and Their Purpose

- **Income**: Annual income of the user.
- **Age**: Age of the user.
- **Experience**: Number of years of professional experience.
- **Married**: Marital status (Married/Single).
- **Ownership**: Car ownership status.
- **Risk flag**: Indicates if the user has defaulted on a loan.

## Key Categorical Attribute

- **Risk Flag**: This attribute is crucial as it directly correlates with the target variable, playing a significant role in predicting loan defaults.

## Attribute with Minimal Impact

- **Marital Status**: Visualization and correlation analysis indicate that marital status has minimal impact on the target variable, showing negligible correlation with loan default.

## Outlier Detection and Class Imbalance

- Outliers were identified using boxplots, but none were found, indicating that the dataset is clean.
- The dataset exhibited class imbalance, which was addressed using Synthetic Minority Over-sampling Technique (SMOTE) to balance the classes.

## Data Encoding

- Categorical data was encoded using one-hot encoding, dropping the first column to avoid multicollinearity.

### Data Scaling

- Only the necessary numerical data was scaled using MinMaxScaler. For instance, age was scaled within the typical human lifespan range (0-80 years).

### Data Transformation

- A function transformer with `np.log1p` was used to transform the data, making it more linear. The Power Transformer was not used due to the presence of zero values in the data.

### Model Selection and Performance

- Initially, logistic regression was chosen for its suitability for classification problems. However, it did not perform well in accuracy tests.
- The decision tree classifier was then employed, which effectively handled the classification problem by creating decision nodes and clusters, resulting in an accuracy score of 87%.

## Conclusion

The decision tree classifier proved to be the most effective model for this classification problem, achieving an accuracy score of 87%. This project demonstrates the power of data analytics and machine learning in accurately predicting loan eligibility based on consumer behavior, thereby assisting financial institutions in making more informed lending decisions.

## Recommendations for Future Work

- **Model Improvement**: Experiment with ensemble methods like Random Forest or Gradient Boosting to potentially improve accuracy.
- **Feature Engineering**: Further exploration of feature engineering techniques to enhance model performance.
- **Hyperparameter Tuning**: Perform extensive hyperparameter tuning for the chosen models to achieve better accuracy.
- **Additional Data**: Incorporate more features and a larger dataset to enhance model generalization and robustness.

By implementing these recommendations, the loan prediction model can be further refined, ensuring higher accuracy and better decision-making capabilities for financial institution