📊 Project Report

Title: ***Predicting Desired Savings Using Socioeconomic Factors***

Objective

To build predictive models that estimate individuals' **desired savings** using a minimal set of socioeconomic indicators, enabling insights into saving behavior for financial planning and personalized services.

Dataset Summary

- **Size**: 20,000 individuals
- **Target Variable**: `Desired_Savings`
- **Input Features Used**:
    1. **Income**
    2. **Age**
    3. **Dependents**
    4. **Occupation** *(OneHotEncoded)*
    5. **City_Tier** *(OrdinalEncoded: Tier_3 < Tier_2 < Tier_1)*
    6. **Expenses**
- **Preprocessing Pipeline**:
    - Missing column (`Unnamed: 16`) dropped
    - Categorical variables encoded with `OrdinalEncoder` and `OneHotEncoder`
    - Standardization applied using `StandardScaler`
    - Dataset split 80/20 for training and testing

🔧 Modeling Pipeline

1. Multiple Linear Regression (MLR)

- **R² Score**: 0.9142
- Trained using the six listed features
- **Diagnostics**:
    - Residuals vs Predicted plots
    - Q-Q Plot and **Shapiro-Wilk Test** → residuals not normally distributed
    - **Breusch-Pagan Test** → heteroscedasticity present
    - **Durbin-Watson** ≈ 2.003 → no autocorrelation
    - **VIF Analysis**: Some multicollinearity found, particularly in encoded categorical variables

2. Regularized Linear Models

- **LassoCV**:
  - Best α = 0.01
  - **R^2 Score**: 0.9142
- **RidgeCV**:
  - Best α = 1.0
  - **R^2 Score**: 0.9142
- **GridSearchCV** was used to fine-tune the regularization parameter `alpha` in both Lasso and Ridge.

3. XGBoost Regressor

- **R^2 Score**: 0.9148
- Parameters: `n_estimators=100`, `learning_rate=0.1`, `max_depth=3`
- Effectively handled:
  - Non-linear patterns
  - Feature interactions
  - Multicollinearity
- Slightly outperformed MLR while being more robust

4. Random Forest Regressor

- **R^2 Score**: **0.96**
- **Best performing model**
- Trained with 100 trees (`n_estimators=100`)
- Provided highest predictive accuracy and generalization capability
- Robust to outliers, feature interactions, and assumptions

Summary Highlights

- Developed a **Multiple Linear Regression model (R^2 = 0.9142)** using features like **income, age, dependents, occupation, city tier, and expenses** to estimate individuals' desired savings.
- Conducted assumption diagnostics including residual plots, Q-Q plot, and VIF analysis, revealing violations of linearity, normality, and multicollinearity.
- Applied **LassoCV and RidgeCV** regularization techniques; leveraged **GridSearchCV** for optimal hyperparameter tuning and feature selection.
- Implemented an **XGBoost Regressor (R^2 = 0.9148)** which matched MLR performance but handled nonlinearities and feature interactions more effectively.
- Trained a **Random Forest Regressor (R^2 = 0.96)** that delivered the best overall accuracy and stability, making it suitable for real-world deployment.

Conclusion

This project demonstrates how a small but meaningful set of socioeconomic features can accurately predict individuals’ desired savings. While linear models provide interpretability, ensemble methods like **Random Forest** and **XGBoost** offer superior performance, especially under assumption violations. The Random Forest model stands out as the most reliable choice for production-level deployment.