**Final Report**

**Costa Rica Household Poverty Prediction**

The Order of the PyTorch

Onur Buyukkalkan, Yi-Huai Chang, Diyanet Nijiati

**Introduction**

The objective of this project is to predict household poverty levels in Costa Rica. The dataset includes a variety of household characteristics that can be used to predict the level of poverty in each household. Accurate predictions are critical for directing aid to those in need. Our approach involves thorough data exploration, cleaning, feature engineering, and testing various machine learning models.

**Data Exploration and Preparation**

**Initial Exploration**

We began by loading the dataset and performing an initial exploration to understand its structure and contents. The dataset contains 143 columns, including both numeric and categorical features. Some columns had significant amounts of missing data, which we needed to address.

**Plotting**

Visualizations played a crucial role in our understanding of the data. We plotted the distribution of the target variable (poverty levels) and used correlation heatmaps to identify relationships between different features. The key finding here is that our target variable is unbalanced and need further adjustment to achieve better prediction. Correlation heatmaps helped us identify redundant features and potential new features that could be useful for our models.

**Data Cleaning**

**Handling Missing Values**

To handle missing values, we filled numeric columns with their median values and dropped columns with more than 70% missing data. This approach ensured that we retained as much information as possible while dealing with incomplete data.

**Feature Conversion**

Categorical text features were converted into numeric values to make them suitable for machine learning algorithms. For example, we converted the "*edjefe*" and "*edjefa*" columns, which indicate the education level of the household head, into dummy variables to indicate whether the household head has years of education larger than 0.

**Feature Engineering**

**Creating New Features**

We generated several new features to better capture the household characteristics:

* **Living Conditions Index:** A composite score based on the quality of household amenities (e.g., wall material, toilet facilities).
* **Economic Stability Opportunity Index:** A score representing the household’s economic stability and opportunities, calculated from assets and education levels.
* **Social Dynamics Index:** A ratio of dependents to adults in the household, reflecting potential financial strain.

These new features were intended to enhance the predictive power of our models by providing additional insights into household conditions.

**Model Testing**

**Data Splitting and Standardization**

We split the data into training, validation, and test sets. The training set was used to train the models, the validation set to tune hyperparameters, and the test set to evaluate final model performance. Standardization was applied to ensure that all features were on a comparable scale.

**Addressing Class Imbalance**

The dataset was imbalanced, with significantly more households in lower poverty levels (non-vulnerable households). To address this, we used Synthetic Minority Over-sampling Technique (SMOTE) to balance the class distribution in the training set.

**Model Selection**

We tested several machine learning models to find the best performer for our task:

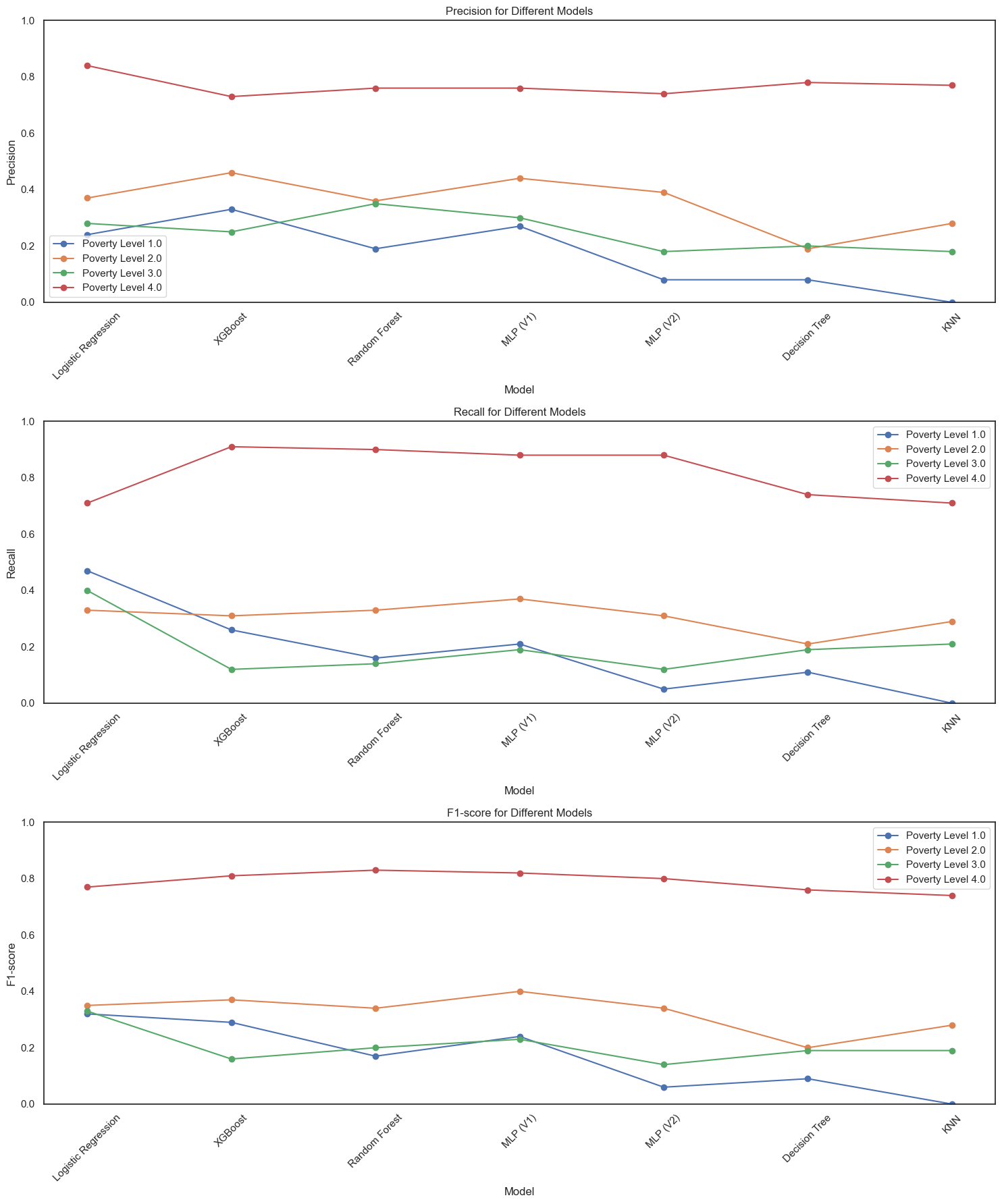
* **Logistic Regression:** Provided a good balance between simplicity and performance, particularly for higher poverty levels.
* **XGBoost:** A powerful gradient boosting model that showed high overall accuracy but struggled with recall for higher poverty levels.
* **Random Forest:** Another ensemble method that performed well but had similar issues with recall.
* **Multi-Layer Perceptron (MLP):** Tested in two versions with varying complexity; did not significantly improve recall rates.
* **Decision Tree:** Optimized for recall but overall performance was lower.
* **K-Nearest Neighbors (KNN):** Showed the lowest performance among the models tested.

**Results and Conclusions**

**Importance of Recall**

In the context of poverty prediction, recall is more important than precision because it is crucial to identify as many households in need as possible. Missing out on households that genuinely need aid (false negatives) is a more significant issue than providing aid to a few non-poor households (false positives).

**Model Performance**

****

* **Logistic Regression:** Achieved the best recall for higher poverty levels, making it the preferred model despite its simplicity.
* **XGBoost and Random Forest:** Showed high accuracy but lower recall, indicating a need for further tuning or additional features to improve recall.
* **MLP, Decision Tree, and KNN:** Did not significantly outperform simpler models and were more computationally intensive.

**Challenges and Limitations**

* **Class Imbalance:** Despite using SMOTE, models had difficulty accurately predicting minority classes.
* **Data Quality:** Inherent noise and missing values in the dataset affected model performance.
* **Model Complexity:** Simpler models sometimes performed better in terms of recall for higher poverty levels.

**Future Work**

* **Advanced Feature Engineering:** Creating more sophisticated features that better capture the nuances of poverty. For instance, generate interaction features that capture the relationships between different variables. Also, Due to the redundancy of some of our variables, using techniques like Principal Component Analysis (PCA) or feature selection algorithms to reduce dimensionality and select the most relevant features may also help increase our accuracy. PCA can help identify the most significant features, reducing noise and improving model performance.
* **Hyperparameter Tuning:** Further fine-tuning the hyperparameters of models, especially neural networks, to improve performance.
* **Ensemble Methods:** Combining predictions from multiple models to leverage their strengths.

**Conclusion**

Our project demonstrated that simpler models like logistic regression could achieve better recall for higher poverty levels, which is critical for identifying households in need. Future work should focus on improving feature engineering, further hyperparameter tuning, and exploring ensemble methods to enhance model performance.