Q1: Missing values were handled separately for numerical and categorical single-response features. For numerical columns, the median was used to impute missing values, reducing the influence of outliers. Categorical responses were filled with "Unknown" to retain information without introducing bias. Ordinal responses such as "I do not know" were encoded as -1 to indicate a distinct, non-informative response while preserving order.

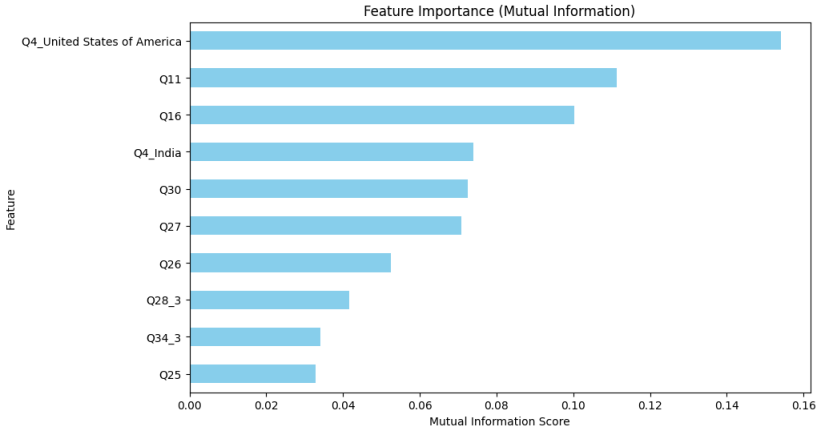
Categorical features were encoded using both ordinal and nominal techniques. Ordinal features, including education level and years of experience, were encoded with increasing integers to maintain order. For high-cardinality nominal features like country, only the top five countries were retained and the rest grouped as “Other.” Remaining nominal features with low cardinality were one-hot encoded to avoid imposing artificial ordering.

Multi-response features (those with underscores in column names) were processed using binary encoding: values were converted to 1 if selected and 0 if missing. This approach preserved all responses while making the data suitable for modeling.

Q2: Feature selection was primarily conducted using Mutual Information (MI), which quantifies how much knowing a feature reduces uncertainty about the target variable. MI is flexible, capturing both linear and nonlinear relationships, and works well with mixed data types, making it ideal for this dataset.

Top features based on MI scores included country (Q4_United States of America), years of coding experience (Q11), years of ML experience (Q16), Q4_India, and ML/AI spending (Q30). These features exhibited the highest information gain with respect to salary predictions.

A new binary feature, “Uses Advanced ML,” was engineered by aggregating deep learning-related tools from multi-response question Q18. This created a practical indicator of technical depth and improved the feature set by capturing specialized skills beyond raw experience.



.

Q3: Cross-validation was performed using a 10-fold approach to assess model performance across different subsets of the training data. The average accuracy obtained across folds was 0.3909, with a variance of 0.0001. The low variance suggests that the model performs consistently across different folds, indicating stability in its generalization ability.

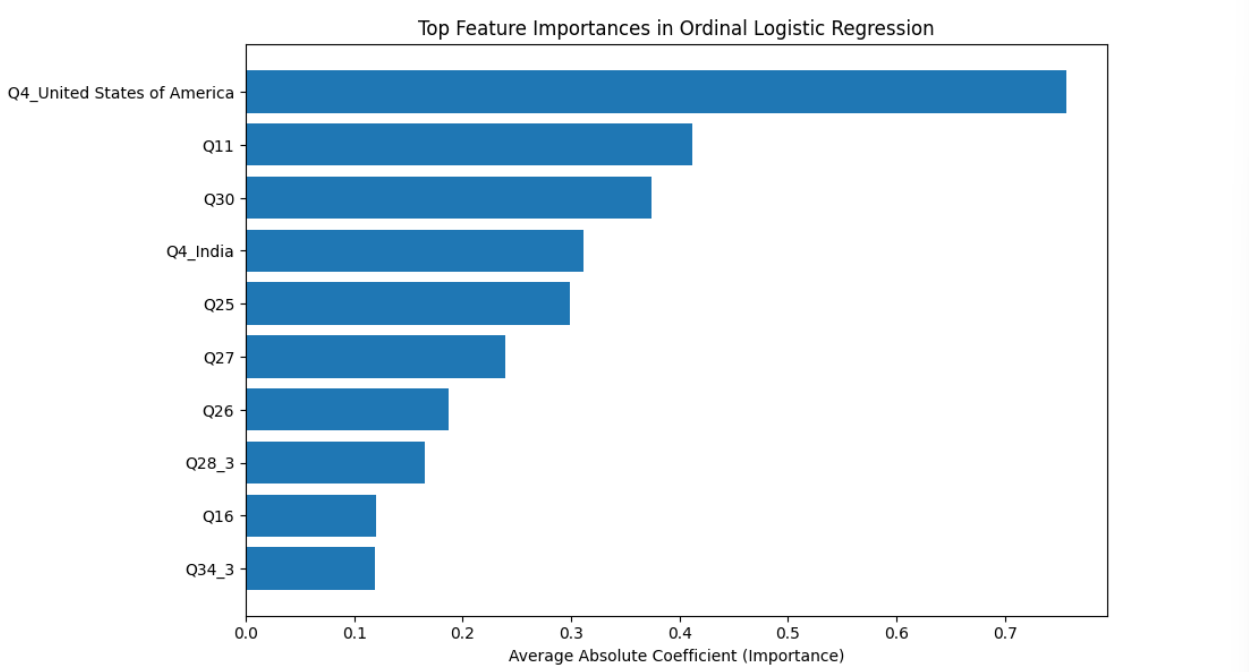
Bias-variance decomposition was conducted to analyze the model's tradeoff between underfitting and overfitting across different values of the regularization parameter C. As expected, higher regularization (C = 0.01) led to higher bias (23.3953) and lower variance (1.0308), indicating underfitting. Conversely, lower regularization (C = 100.0) resulted in reduced bias (22.5561) but slightly higher variance (1.2644), signaling a slight increase in overfitting. The most balanced model was observed around C = 0.1, where bias (22.5607) and variance (1.2802) jointly achieved the best tradeoff.

Feature scaling was incorporated prior to model fitting to ensure comparability of coefficients across both ordinal and one-hot encoded variables, improving interpretability of feature importance.

Q4: Accuracy was not chosen as the primary evaluation metric because it does not account for the ordinal nature of the salary categories. Misclassifying a low-salary category as a high-salary category (or vice versa) carries more significance than minor misclassifications between adjacent categories. Instead, Quadratic Weighted Kappa (QWK) was selected as it penalizes larger misclassification errors more heavily and measures agreement between predicted and actual salary levels while considering ordinal relationships.

Hyperparameter tuning was performed using Grid Search with 5-fold cross-validation, optimizing for QWK. The key hyperparameters tuned were regularization strength (C) and max iterations (max_iter). The best-performing model had C = 1 and max_iter = 100, achieving a QWK score of 0.4381. Class balancing techniques were not applied, which may have contributed to poor prediction of underrepresented salary classes.

Feature importance was analyzed using the average absolute coefficient values from the k−1 binary models. The most influential predictors included country (Q4_United States of America, Q4_India), years of coding experience (Q11), money spent on ML (Q30), and company size (Q26). These findings show a partial overlap with the MI-based rankings but highlight stronger model reliance on demographic and investment-related factors rather than just technical experience. Practical ML engagement remained relevant, but geographic and organizational context emerged as stronger salary indicators in the final model

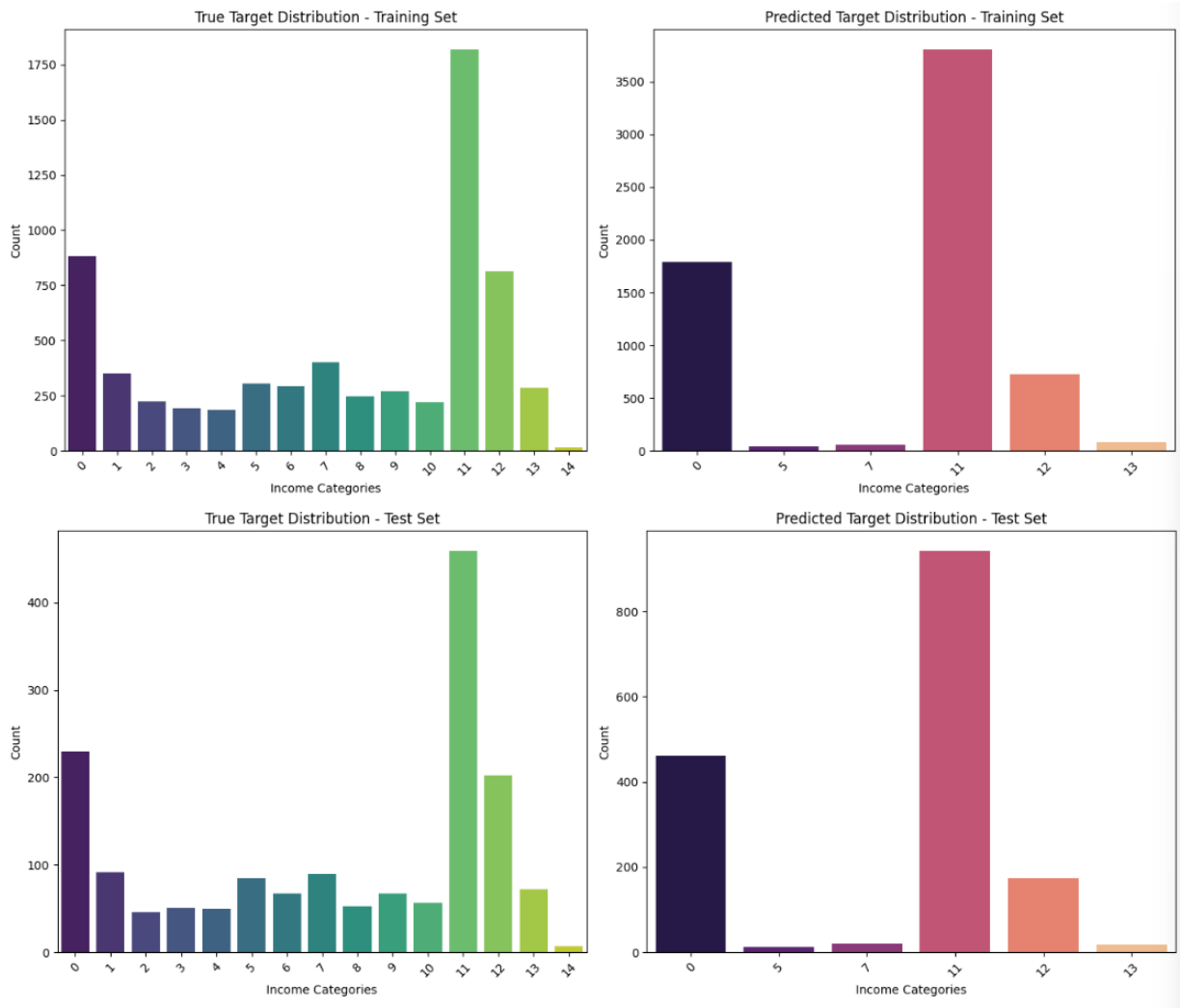


Q5: The Quadratic Weighted Kappa (QWK) score was used to evaluate model performance on both training and test sets. The training set QWK was 0.4813, while the test set QWK was 0.4782, indicating minimal performance drop on unseen data. Similarly, accuracy remained consistent across datasets, with 0.3906 on training and 0.3821 on test. These small differences (QWK gap = 0.0031, accuracy gap = 0.0085) suggest the model achieved a strong bias-variance tradeoff and generalized well without significant overfitting or underfitting.

Overfitting analysis confirmed this conclusion. The minimal performance gap indicated that the model learned meaningful patterns from the training data while maintaining robustness when evaluated on unseen samples. Hyperparameter tuning (e.g., regularization strength C = 1.0) helped balance model complexity and generalization. Therefore, further regularization adjustments were not immediately necessary based on current metrics.

Despite good quantitative performance, the distribution plots highlighted a challenge. The model tended to favor a few dominant salary categories, particularly low and high extremes, while underpredicting less frequent mid-range income brackets. Although this behavior is partially explained by class imbalance in the dataset, it suggests room for improvement in recall across underrepresented classes.

Overall, the model successfully captured the general structure of the salary distribution and prioritized interpretable features like experience, country, and ML usage. However, performance could be further enhanced by exploring techniques to address class imbalance, such as cost-sensitive learning or oversampling underrepresented salary ranges during training.


True Target Distribution - Training Set
Count
Income Categories
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14
0 250 500 750 1000 1250 1500 1750
Predicted Target Distribution - Training Set
Count
Income Categories
0 5 7 11 12 13
0 500 1000 1500 2000 2500 3000 3500
True Target Distribution - Test Set
Count
Income Categories
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14
0 100 200 300 400
Predicted Target Distribution - Test Set
Count
Income Categories
0 5 7 11 12 13
0 200 400 600 800