Introduction

This report presents an analysis of the CPS-Earnings dataset, focusing on predicting hourly earnings for the occupation **Retail Salespersons** (coded as **4760.0**) using linear regression models of increasing complexity. The primary objective is to evaluate model performance using RMSE, cross-validated RMSE, and BIC metrics to select the most suitable predictive model.

Data Preparation and Occupation Selection

The CPS-Earnings dataset was loaded and filtered to include only individuals working in the occupation **Retail Salespersons**. This occupation was chosen due to its robust sample size, ensuring statistical reliability. The data was cleaned by removing missing values for critical variables such as hourly earnings (**earnhre**). Categorical variables, including **sex**, **race**, **marital status**, **union membership**, **ethnicity**, and **presence of children**, were converted into dummy variables to facilitate their inclusion in linear regression models.

Predictive Model Design

Four linear regression models were developed with increasing complexity:

- **Model 1:** Included basic predictors such as **age**, **education level** (**grade92**), and **sex**.
- **Model 2:** Added **race** variables to capture demographic influences.
- **Model 3:** Introduced **marital status** and **union membership** variables to account for socioeconomic factors.
- **Model 4:** Added **ethnicity**, and **presence of children** to incorporate household characteristics.

Model Performance Summary

The table below provides an overview of the model selection process, including the number of variables and coefficients, RMSE, and BIC for each model.

Table 1: Breakdown of Model Performance

Model	N vars	RMSE	CV RSME	BIC
Model 1	3	374.49	449.64	30812.76
Model 2	4	374.60	450.34	30940.15
Model 3	6	379.12	445.88	31057.54
Model 4	8	383.49	446.30	31289.97

Conclusion

The analysis indicates a trade-off between model simplicity and generalization performance:

Model 1 offers the best balance of simplicity and interpretability, achieving the lowest **RMSE** (374.49) and **BIC** (30812.76). This makes it the preferred model when prioritizing in-sample fit and minimizing model complexity. However, **Model 3** demonstrates the lowest **CV RMSE** (445.88), suggesting stronger generalization performance to unseen data. While this model incorporates additional socioeconomic variables, it maintains a reasonable complexity and avoids the overfitting seen in Model 4. Overall, the analysis highlights the importance of balancing model complexity with predictive performance, ensuring robust and generalizable insights.

Appendix

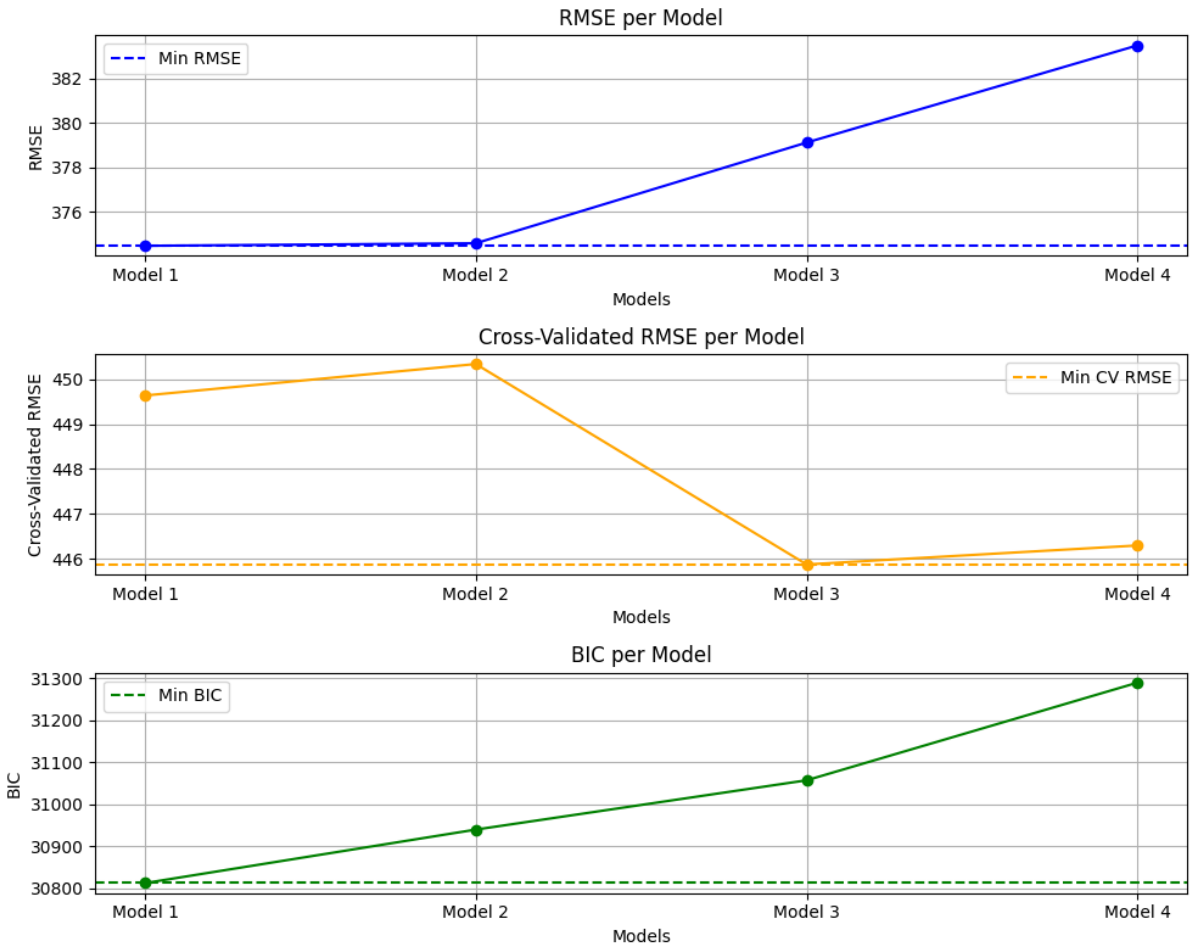

Figure 1: Model BIC, RSME, and CV RSME Comparison

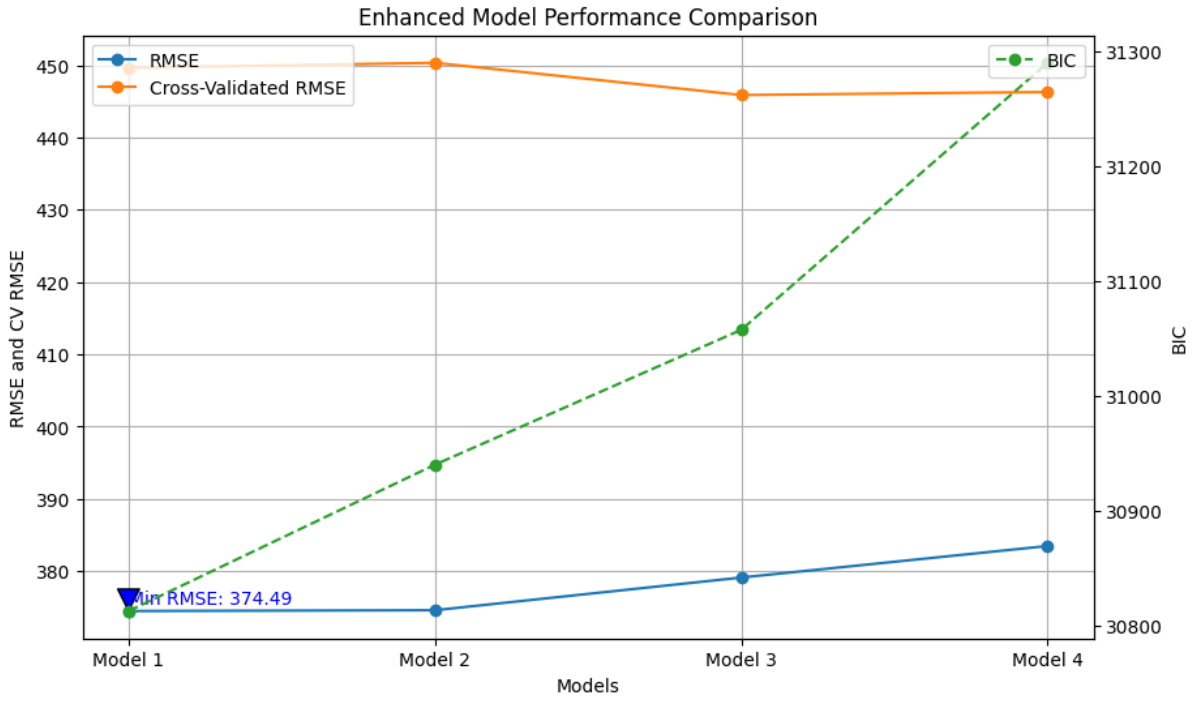

Figure 2: Enhanced Model Performance Comparison