

## Phase 2 Task 5: Recommendations

*Team 2: Tsyr Rau Chen, Arailym Duisengali, Sheikh Noohery, Lo Ying Wu, Kuan Rong Yang*

We recommend the 2nd degree polynomial regression model for predicting tipping behavior in the food delivery dataset. However, it is important to note that all models tested show significant limitations and poor overall performance that we will discuss in the end. This model has the best balance between accuracy and simplicity, with a Root Mean Square Error (RMSE) of ~7.99 and Mean Absolute Error (MAE) of ~6.22. The RMSE value for the 2nd degree polynomial regression model is the lowest among all models tested. Even though the RMSE value is slightly better than the linear regression model, this small improvement suggests that the 2nd degree polynomial model is better at explaining the variability in tip amounts.

The residual plots for the models explain why the 2nd degree polynomial performs best. The simple linear regression shows distinct horizontal banding patterns in its residuals, indicating that the model misses important non-linear relationships in the data. The 2nd degree polynomial model's residual plot displays more random scatter around zero. The 3rd and 4th degree polynomial regression models show increasingly erratic residuals with extreme outliers. This behavior makes sense since complex models have a risk of overfitting the training data. The 3rd and 4th degree polynomial models are less reliable for predicting tips for new deliveries.

Surprisingly, the random forest model performs significantly worse than all of the polynomial models, with RMSE of ~8.85 and MAE of 6.83. This unexpected result suggests that our dataset does not work well with tree-based models. The reasoning could be that the relationships are smooth and continuous rather than step-wise patterns that decision trees capture the best.

Despite our selection of the 2nd degree polynomial model as the best model, all models show significant limitations that indicate poor predictive performance. An RMSE of ~8.0 means the average prediction error is substantial compared to typical tip amounts in the dataset. Additionally, all models show inconsistent error patterns and outliers in their residual plots, meaning the models struggle to make accurate predictions across different tip ranges. Also, a sophisticated model like Random Forest not outperforming simple linear regression indicates that our predictor variables (distance, duration, and city) don't provide enough information to accurately predict tip amounts. Tipping behavior is likely affected by additional factors such as delivery timeliness, weather conditions, or order value. Maybe a better approach to modeling could be predicting tip categories, like low, medium, or high, instead of exact dollar amounts.