Predicting Learning Commons Usage: Duration and Occupancy

A Statistical Learning Approach

Emma Naiyue Liang, Ryan Rankin, Jaryt Salvo, & Jason Turk

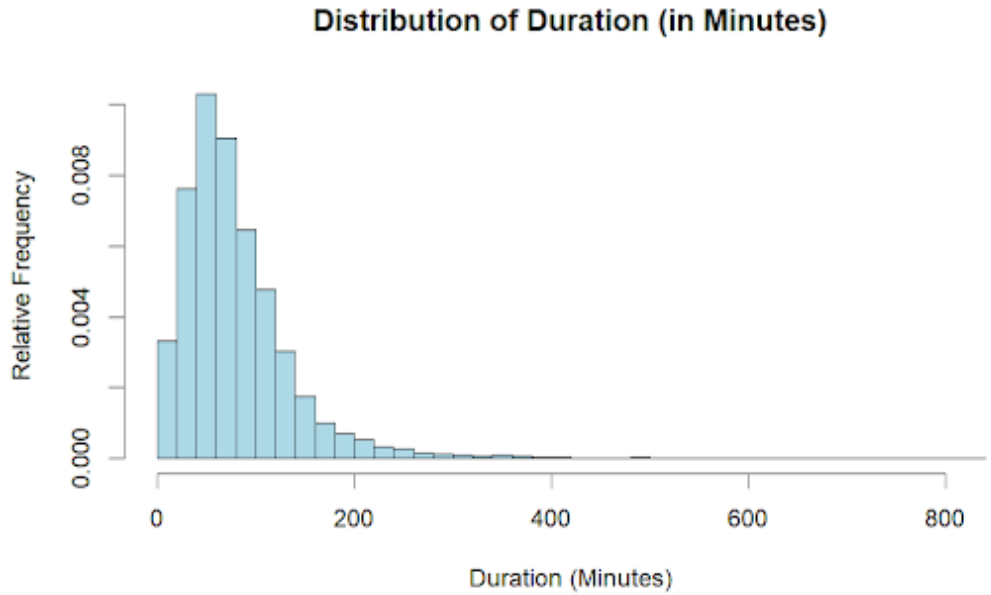
MATH 7550 Statistical Learning || BGSU

December 4, 2024

Distribution Statistics

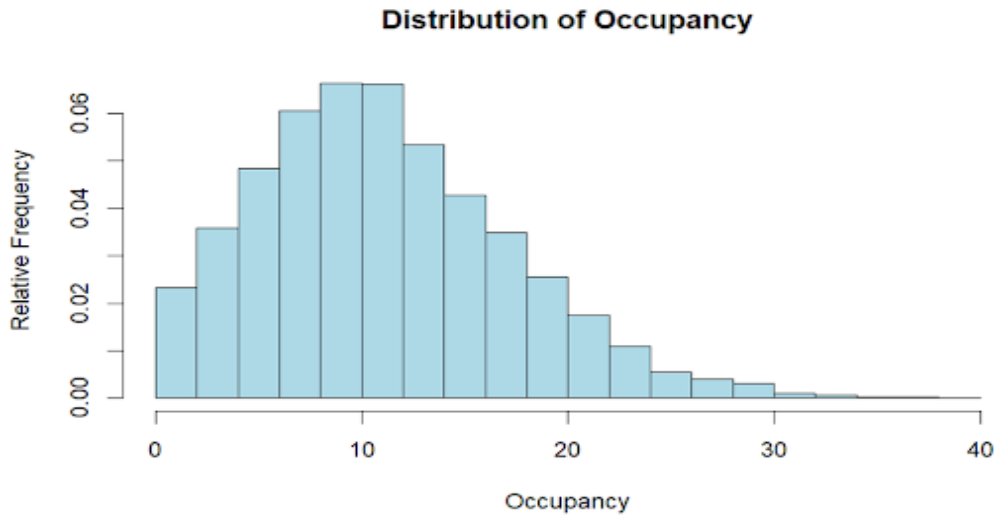
Duration (minutes):

Statistic	Value
Minimum	6.00
1st Quartile	44.00
Median	68.00
Mean	81.78
3rd Quartile	103.00
Maximum	822.00



Occupancy (students):

Statistic	Value
Minimum	1.00
1st Quartile	7.00
Median	11.00
Mean	11.62
3rd Quartile	15.00
Maximum	40.00



Feature Categories Overview

Category	Key Features
Temporal	Time of day, Day of week, Week of semester
Academic	Course level, GPA categories, Credit load
Visit	Duration patterns, Group sizes, Visit frequency
Course	Subject areas, Level progression, Course mix
Student	Major groups, Class standing, Academic progress

Engineering Approach

- Temporal patterns
- Academic context
- Student behavior
- Group dynamics

Dropped Raw Features

Raw Feature	Engineered Into
Student_IDs	Total_Visits, Semester_Visits, Avg_Weekly_Visits
Class_Standing	Class_Standing_Self_Reported, Class_Standing_BGSU
Major	Major_Category, Has_Multiple_Majors
Expected_Graduation	Expected_Graduation_Date, Months_Until_Graduation
Course_Name	Course_Name_Category
Course_Number	Unique_Courses, Course_Level_Mix
Course_Type	Course_Type_Category
Course_Code_by_Thousands	Course_Level, Advanced_Course_Ratio

Feature Engineering Strategy

Raw features were transformed into more informative derived features, capturing higher-level patterns and relationships in the data.

Complete Feature List (~50 Pre-Dummied)

Category	Features
Student Demographics	**Student_IDs**, Gender, **Class_Standing**, Class_Standing_Self_Reported, Class_Standing_BGSU, Has_Multiple_Majors, **Major**, Major_Category, Degree_Type
Academic Performance	**Total_Credit_Hours_Earned**, Term_Credit_Hours, Credit_Load_Category, Term_GPA, Cumulative_GPA, Change_in_GPA, GPA_Category, GPA_Trend
Course Information	**Course_Name**, **Course_Number**, **Course_Type**, Course_Type_Category, Course_Level, **Course_Code_by_Thousands**, Course_Name_Category, Course_Level_Mix, Advanced_Course_Ratio, Unique_Courses
Temporal Features	**Check_In_Time**, **Check_Out_Time**, **Check_In_Date**, Check_In_Hour, Check_In_Day, Check_In_Month, **Semester**, Semester_Week, Check_In_Week, Is_Weekend, Time_Category
Visit Metrics	**Duration_In_Min**, Group_Size, Group_Size_Category, Group_Check_In, Total_Visits, Semester_Visits, Week_Volume, Avg_Weekly_Visits, **Occupancy**
Graduation	**Expected_Graduation**, Expected_Graduation_Date, Months_Until_Graduation

Common Models for Both Tasks

Model	Hyperparameters
Ridge	$\alpha \in [10^0, 10^2]$
Lasso	$\alpha \in [10^{-2}, 10^0]$
Penalized-Splines	**knots**: $\{9, 11, 13, 15\}$, **degree**: 3, **ridge**: $\alpha \in [10^0, 10^2]$
KNN	**neighbors**: $\{15, 17, 19, 21\}$, **weights**: {uniform, distance}

Implementation Details

- **Duration**: Log-normal
- **Occupancy**: Poisson & Weibull
- Integer rounding for occupancy
- Grid search optimization
- Feature selection
- Cross-validation

Model Pipeline Configurations

CV Method	Description
kfold	Random k splits
rolling	Fixed-size window moving forward
expanding	Growing window with fixed start point

Pipeline	Implementation
vanilla	Scaling → Model
interact_select	Scaling → Interactions → SelectKBest → Model
pca_lda	Scaling → PCA/LDA → Interactions → SelectKBest → Model

Scaling Methods

- **StandardScaler**: $(x - \mu)/\sigma$ - sensitive to outliers
- **RobustScaler**: $(x - Q_2)/(Q_3 - Q_1)$ - resistant to outliers
- **MinMaxScaler**: $(x - x_{min})/(x_{max} - x_{min})$ - preserves zeros

Best Model Configurations

Duration Prediction:

Component	Value
Model	PenalizedSplines
Pipeline	vanilla
CV Method	kfold
RMSE	59.47
R^2	0.059
Ridge α	14.38
Spline degree	3
Spline knots	15
Scaler	RobustScaler

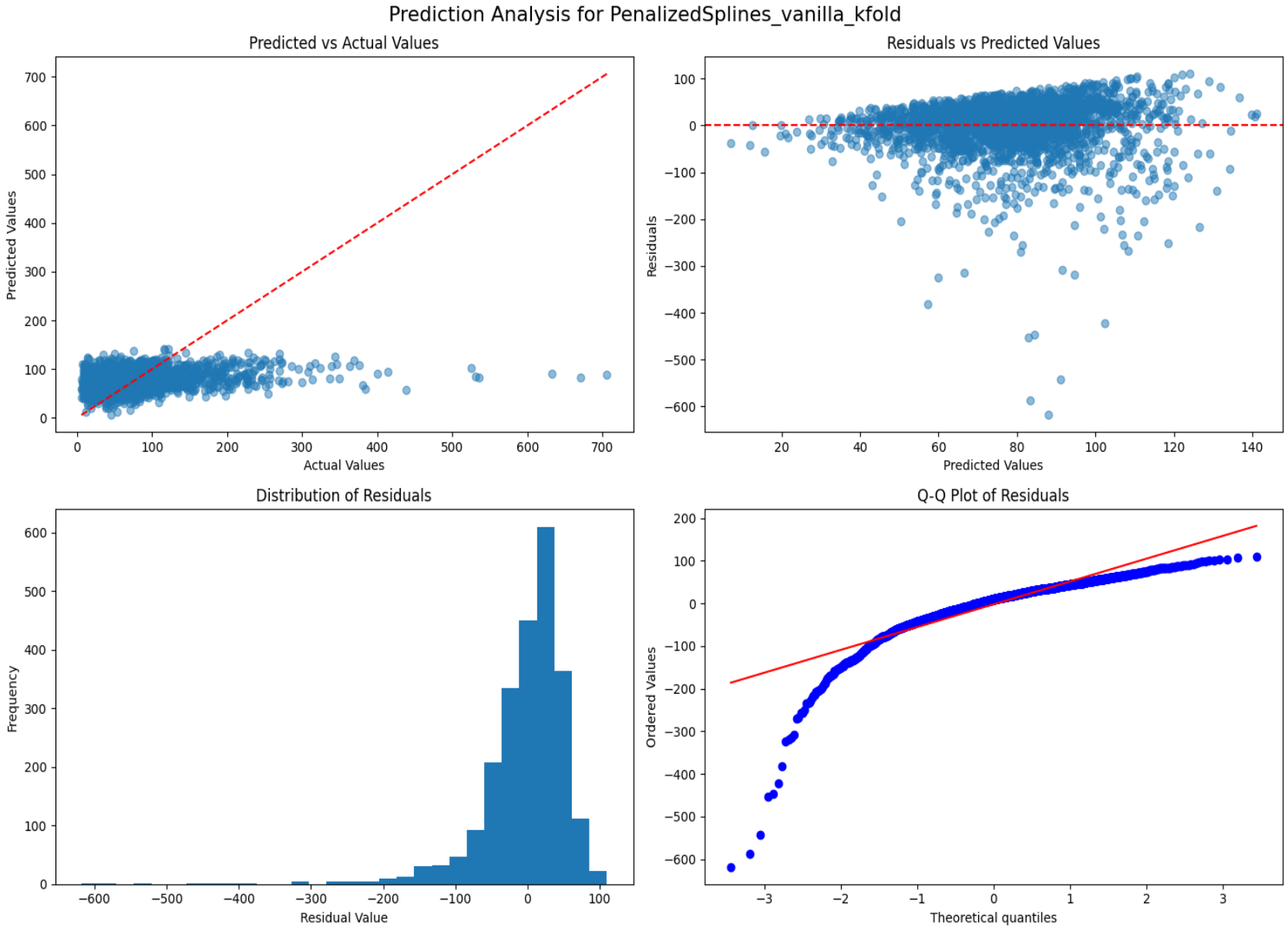
Occupancy Prediction:

Component	Value
Model	PenalizedSplines
Pipeline	vanilla
CV Method	rolling
RMSE	3.64
R^2	0.303
Ridge α	29.76
Spline degree	3
Spline knots	15
Scaler	RobustScaler

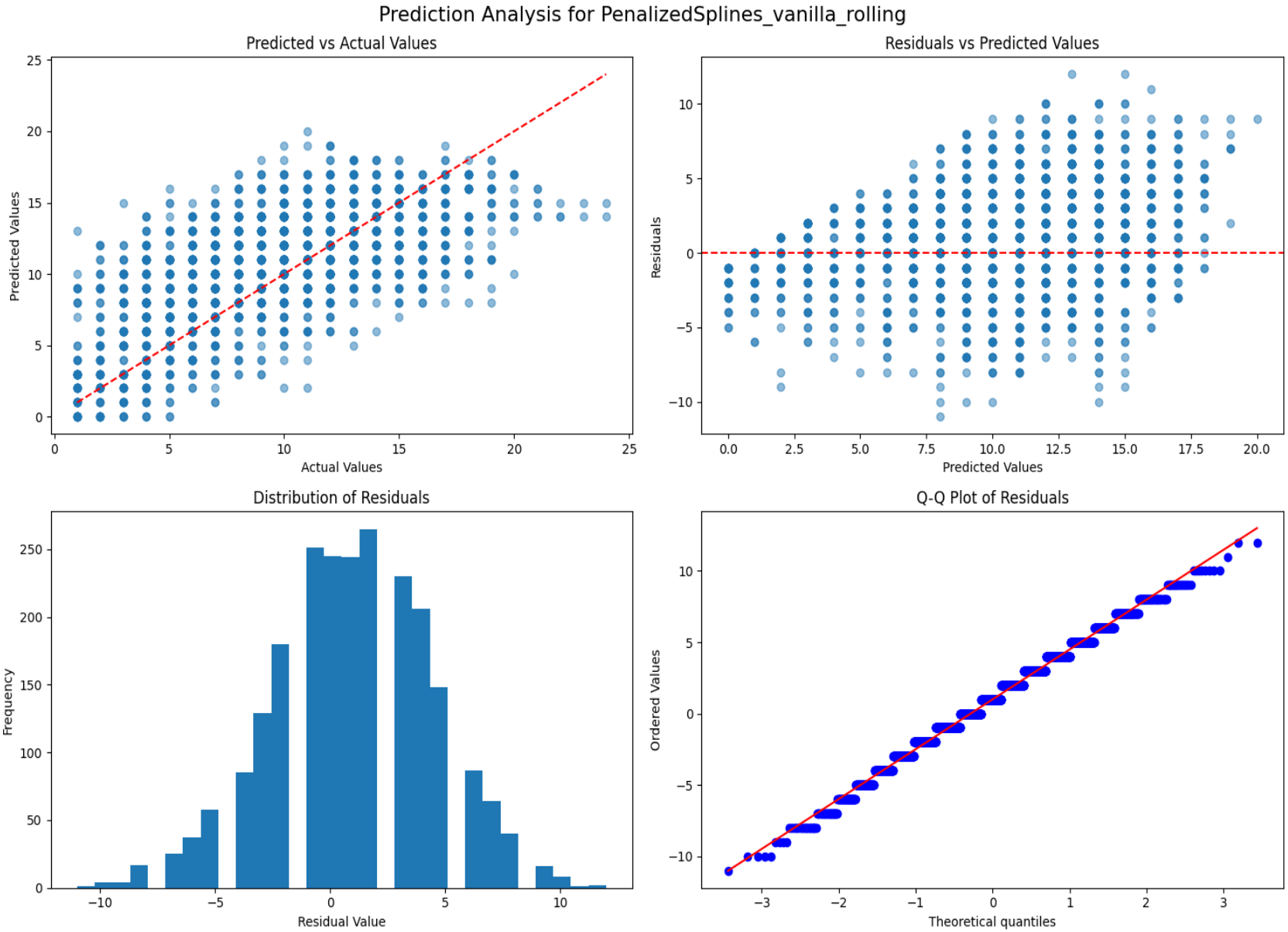
Key Insight

Both tasks achieved best results with PenalizedSplines and vanilla features, though with different CV methods & regularization.

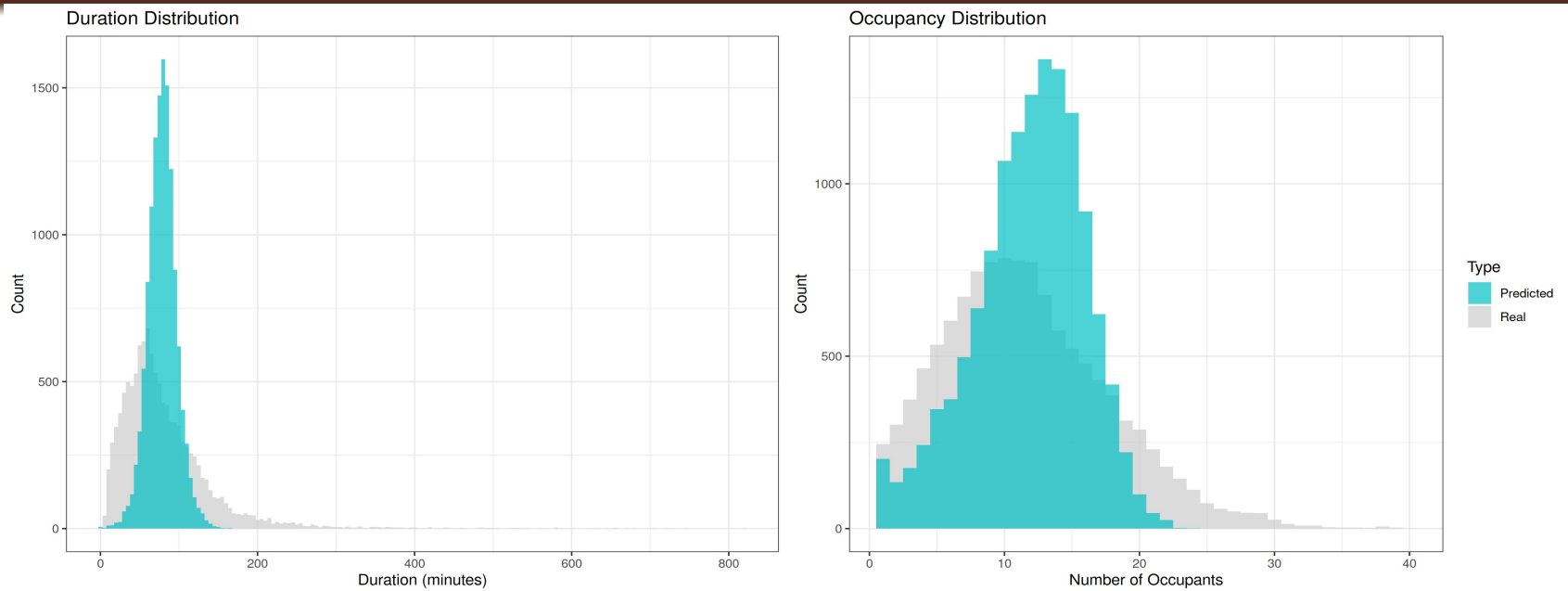
Duration: Best Model Diagnostics

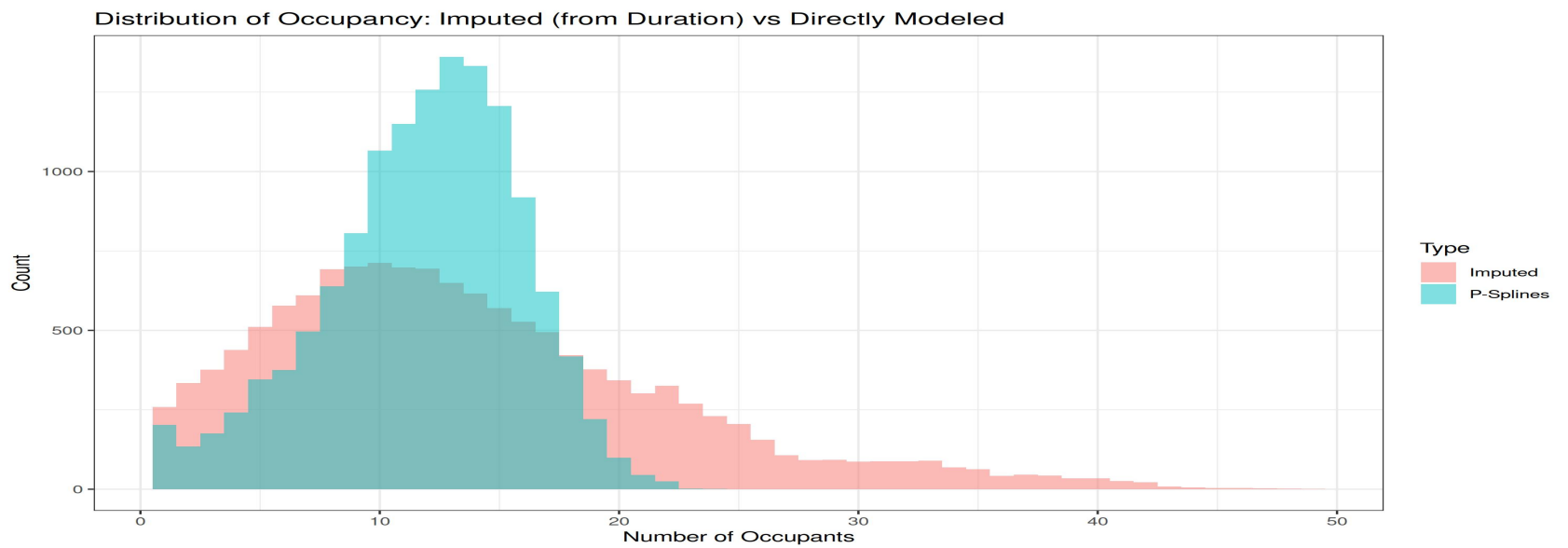


Occupancy: Best Model Diagnostics



Sanity Check





Key Findings

Main Results:

- **PenalizedSplines** with *vanilla* features performed best
- **Occupancy** prediction shows promise ($R^2 = 0.303$)
- **Duration** prediction remains challenging ($R^2 = 0.059$)

Future Directions:

- Incorporate *weather* data
- Explore *non-linear* relationships further
- Investigate *time series* approaches

Impact

While duration prediction remains difficult, our occupancy model shows strong potential for a victory **#CautiousOptimism**

Thank You

For Your Attention

Questions & Discussion Welcome