Predicting Learning Commons Usage: Duration and Occupancy

A Non-Linear Approach

Naiyue Liang (Emma), Ryan Renken, Jaryt Salvo, & Jason Turk

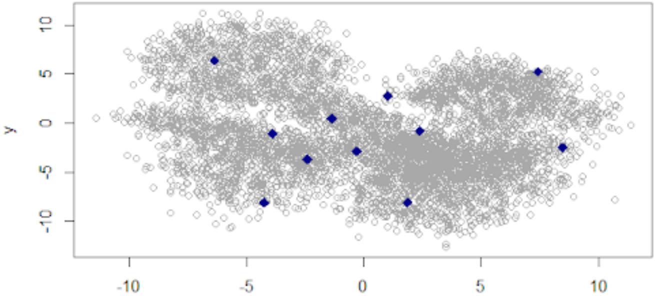
MATH 7560 Statistical Learning II || BGSU

April 24, 2025

K-means: Data & Motivation

Clustering Task

- **Goal**: Partition $\mathbb{R}^2$ data into $k = 11$ clusters.
- **Comparison**: Standard K-means vs K-means++ initialization.
- **K-means++ Motivation**: Better initial centroids for potentially faster/better clustering.



K-means vs K-means++ Results

Standard K-means:

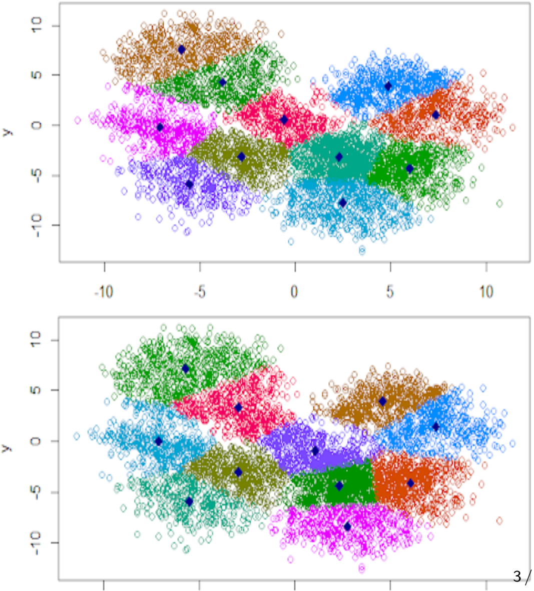
- **Stats**: 5 iterations, WCSS = 22,824.

K-means++ Initialization:

- **Stats**: 8 iterations, WCSS = 22,943.

Observation

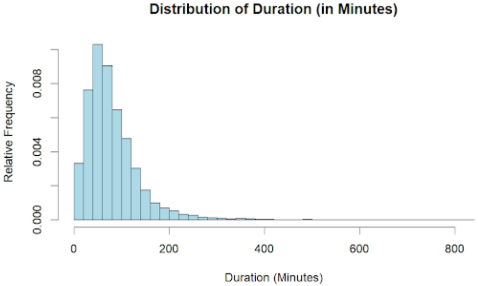
K-means++ offered no clear advantage over standard K-means for this dataset (visually or by WCSS).



Distribution Statistics

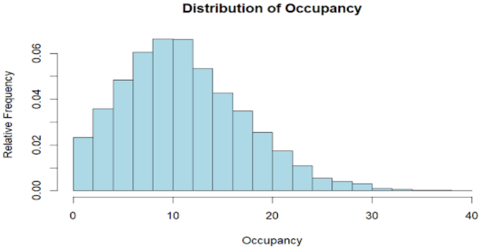
Duration (minutes):

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

External Source	**Key Features**
R library ‘lunar‘	Moon phase data
R library ‘openmeteo‘	Hourly weather metrics (temperature, humidity, pressure, cloud cover, wind, radiation, precipitation, & soil conditions)

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.

Model Hyperparameter Tuning Ranges

Duration Task Models

Model	**Hyperparameters**
MARS	**num_terms**: $[7, 15]$ **prod_degree**: 1
Random Forest	**trees**: $[300, 325]$ **min_n**: $[15, 25]$ **mtry**: $[20, 25]$
XGBoost	**trees**: $[75, 100]$ **tree_depth**: $[15, 21]$ **learn_rate**: 0.05 **min_n**: $[10, 15]$ **mtry**: $[12, 15]$

Occupancy Task Models

Model	**Hyperparameters**
MARS	**num_terms**: $[120, 130]$ **prod_degree**: 1
Random Forest	**trees**: $[250, 350]$ **min_n**: $[2, 3]$ **mtry**: $[40, 45]$
XGBoost	**trees**: $[350, 450]$ **tree_depth**: $[6, 8]$ **learn_rate**: 0.1 **min_n**: $[2, 3]$ **mtry**: $[30, 35]$

Deep Learning Hyperparameters

Duration Task Models

Model	Hyperparameters
GRU	**lr**: $[10^{-3}, 5 \times 10^{-3}]$ **batch_size**: $\{64, 128\}$ **gru_dim**: $\{128, 256, 512\}$ **num_layers**: $\{1, 2\}$ **gru_expansion**: $[0.5, 1.4]$ **dropout_rate**: $[0.25, 0.42]$ **weight_decay**: $[10^{-6}, 10^{-5}]$ **activation_fn**: relu
Trans-former	**lr**: $[10^{-3}, 7 \times 10^{-3}]$ **batch_size**: 64 **d_model**: 128 **nhead**: $\{4, 8\}$ **nlayers**: $\{2, 3\}$ **d_hid**: 128 **dropout**: $[0.05, 0.20]$ **weight_decay**: $[10^{-6}, 10^{-4}]$

Occupancy Task Models

Model	Hyperparameters
GRU	**lr**: $[1.2 \times 10^{-3}, 5 \times 10^{-3}]$ **batch_size**: $\{64, 128\}$ **gru_dim**: 512 **num_layers**: 2 **gru_expansion**: $[0.6, 1.4]$ **dropout_rate**: $[0.25, 0.42]$ **weight_decay**: $[10^{-6}, 5 \times 10^{-6}]$ **activation_fn**: relu
Trans-former	**lr**: $[8 \times 10^{-4}, 1.5 \times 10^{-3}]$ **batch_size**: $\{64, 128\}$ **d_model**: $\{64, 128, 256\}$ **nhead**: $\{4, 8\}$ **nlayers**: $\{2, 3\}$ **d_hid**: 128 **dropout**: $[0.05, 0.20]$ **weight_decay**: $[10^{-6}, 10^{-5}]$

Preprocessing Pipeline Details

R ‘recipes‘ Pipeline Steps

1. Define roles (outcome, predictors)
2. Remove specified ID/date/unwanted columns
3. Convert `Check_In_Time` to minutes past midnight
4. Impute missing numerics (mean)
5. Handle novel factor levels
6. Create dummy variables (drop first)
7. Remove zero-variance predictors
8. Normalize numeric predictors

Step 5: Handling Novel Factor Levels

Prepares for unseen categories in new data:

- Adds a special "novel" level to factors.
- Replaces unknown categories with "novel" instead of causing errors.
- Ensures robust predictions, especially before dummy encoding.

Top Model Performance (Holdout Set)

Duration Task

Model	RMSE	R^2
XGBoost	59.9	0.099
Random Forest	60.1	0.090
Transformer	61.3	0.010
MARS	61.6	0.045
GRU	63.8	0.041

Occupancy Task

Model	RMSE	R^2
XGBoost	1.83	0.911
Random Forest	1.93	0.902
GRU	3.16	0.738
Transformer	3.26	0.706
MARS	3.78	0.617

Key Performance Observations

Based on holdout set metrics:

- **Duration Task:** XGBoost achieved the lowest RMSE (59.9) and highest R^2 (0.099).

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