Hackathon 2024 | IML (67577)| HUJI

Project Description

Chosen Challenge

HU.BER – Optimizing Public Transportation Routes

Authors

Yarden Ashuach 206532822

Omer Dahan 315466664

Yotam Megged 319134912

Shalev Shitrit 314825597

עיבוד מקדים

בשלב העיבוד המקדים, בוצעו מספר שלבים מרכזיים:

1. **מחיקת שורות עם ערכים לא תקינים**

א. שורות ללא זמן הגעה או עם זמן הגעה בפורמט לא תקין.

ב. שורות עם מספר שלילי של אנשים שעולים או ממשיכים.

ג. שורות עם נתוני מיקום (קו אורך ורוחב) לא תקינים - אם סטיית התקן של מיקום מסוים גדולה מדי, כלומר מיקום נמצא באזור מרוחק מאוד מהאזור הממוצע.

2. **מחיקת טורים שהערכים בהם לא רלוונטיים לחיזוי**

א. **trip_id, trip_id_unique, trip_id_unique_station** - ערכים המציגים מספרים מזהים, ללא ערך כמותי. כמו כן, חלק מהמספרים המזהים אף מכילים מזהים אחרים.

ב. **alternative ,part** - ערכים לא מוגדרים היטב.

ג. **station_name** - ערך לא מספרי. ניתן להמיר לקטגוריאלי, אך המידע על התחנה קיים גם במיקום ולכן הוחלט להשמיטו.

3. **שינוי תאים קיימים, יצירת טורים חדשים ומילוי הערכים בהם**

א. **direction** הוחלף לערכים קטגוריאליים של 0 או 1.

ב. **cluster** - הוחלף למשתנים קטגוריאליים באמצעות fit_transform של sklearn.

ג. **is_estimated** הוחלף לערכים קטגוריאליים של 0 או 1.

ד. **time_in_station** - החלפה של door_closing_time בטור זה, שמחושב כזמן סגירת הדלת פחות זמן ההגעה לתחנה. הטור החדש מכיל מידע שעשוי להשפיע על מספר הנוסעים שעולים.

4. **מילוי ערכים חסרים הניתנים לחיזוי**

א. מילוי **ערכים ב- time_in_station** - במקרה שלא היו נתונים על זמן סגירת הדלת (מקרים בהם זמן ההגעה היה משוער) מילאנו את הזמן הממוצע שהאוטובוס חיכה בתחנה.

בחירת מודל

כדי לבחור מודל מתאים, בדקנו האם הנתונים מתנהגים בצורה לינארית. עבור כל פיצ'ר בדקנו את הקורלציה בין הפיצ'ר לבין הלייבל (מספר הנוסעים שעלו או זמן המסלול).

בחלק הראשון, מכיוון שנמצאה קורלציה לינארית עבור פיצ'רים מסוימים, החלטנו להשתמש במודל של רגרסיה לינארית. בנוסף, בדקנו גם מודלים של עץ החלטה ו-Random Forest, מכיוון שלפיצ'רים מסוימים לא הייתה קורלציה לינארית גבוהה והנתונים התאימו גם למודלים אלו.

בחלק השני, בחרנו להשתמש בעץ החלטה במקום ברגרסיה לינארית מכיוון שעץ החלטה מתאים יותר לטיפול בקשרים לא לינאריים ובמשתנים קטגוריים, שנפוצים בעולם התחבורה הציבורית. עץ החלטה מסוגל להתמודד עם ערכים חסרים ולספק תובנות ברורות לגבי הגורמים המשפיעים על התוצאה, בעוד שרגרסיה לינארית מוגבלת בטיפול בקשרים מורכבים ורגישה יותר לערכים חריגים וחסרים.

מסקנות והמלצות

בהתבסס על ניתוח הנתונים והמסקנות שהתקבלו, ניתן להסיק מספר מסקנות בנוגע למערכת התחבורה הציבורית:

1. קווי אוטובוסים עמוסים במיוחד

הקווים העמוסים ביותר שזוהו הם:

קו 311411 (אזור דרומי בת ים).

קו 310990 (אזור חולון עירוני ומטרופוליני + תחרות חולון).

קו 311928 (דרומי-ראשל"צ-חולון).

קווים אלו נמצאים באזורים הסמוכים בת-ים, חולון וראשון לציון, ועוברים באזורים צפופים ומאוכלסים מאוד, אשר ידועים בשימוש רב בתחבורה ציבורית.

המלצות: מומלץ להגדיל את תדירות הקווים באזורים אלו כדי לשפר את השירות ולמנוע עומסים:

- קו 311411 - תדירות הקו צריכה להיות מוגברת בשעות השיא (7:00-9:00 ו-16:00-18:00).
- קו 310990 - תדירות הקו צריכה להיות מוגברת במהלך כל היום.
- קו 311928 - תדירות הקו צריכה להיות מוגברת בשעות השיא (7:00-9:00 ו-16:00-18:00).

בנוסף, ניתן לשקול הוספת קווים תומכים שיפעלו במקביל לקווים אלו ויפחיתו את העומס עליהם.

2. קווים בעלי תפוקה נמוכה

ישנם קווי אוטובוסים בהם יש תפוקה נמוכה לאורך כל היום, דבר המעיד על ביקוש נמוך לשירותיהם. קווים אלו עשויים לעבור באזורים פחות מאוכלסים או באזורים שבהם השימוש בתחבורה ציבורית נמוך בשל זמינות גבוהה של רכבים פרטיים או אמצעי תחבורה חלופיים.

בנוסף, התוואי של קווים אלו עשוי להיות פחות נוח או יעיל לנוסעים, או שהתדירות ושעות הפעילות שלהם אינן מספקות, מה שמוביל למיעוט משתמשים לאורך כל שעות היום.

המלצות: במידה והביקוש לקווים אלו נמוך במיוחד, ניתן לשקול הפחתת התדירות שלהם, בעיקר בשעות השפל, כדי להתאים את השירות לביקוש האמיתי ולחסוך בעלויות תפעול, לדוגמא:

- קו 310567 - תדירות הקו יכולה להיות מופחתת בשעות הבוקר המאוחרות והצהריים.
- קו 311234 - תדירות הקו יכולה להיות מופחתת במהלך כל היום.

כמו כן, מומלץ לבחון את מסלול הקווים ושינויו כדי לשרת אזורים בהם יש ביקוש גבוה יותר. למשל, חיבור הקווים לאזורים מסחריים, מוסדות חינוך או תחנות מרכזיות יותר עשוי להגדיל את מספר הנוסעים.

3. התפלגות עליית נוסעים לאוטובוסים לפי אזורים

ניתוח האזורים השונים הראה כי ישנם אזורים בהם התפלגות העולים על האוטובוסים שונה במהלך היום. לדוגמה:

אזור 6 (בני ברק) - ישנה עלייה גבוהה יחסית בשעות אחר הצהריים.

אזור 9 (אונו/אלעד) - ישנה תפוקה נמוכה לאורך כל היום.

המלצות: באזורים עם עלייה גבוהה בשעות אחר הצהריים, למשל בבני ברק, יכולה הוספת אוטובוסים להגדיל את התדירות בשעות אלו ולספק שירות טוב יותר בזמנים עמוסים. לדוגמא: קו חדש שיפעל בשעות אחר הצהריים ויעבור בתחנות מרכזיות בבני ברק, כדי להקל על העומס בשעות אלו.

בנוסף, באזורים כמו אונו/אלעד, שבהם נרשמת תפוקה נמוכה לאורך היום, מומלץ לשקול ירידה בתדירות האוטובוסים או לאפשר תחבורה פחות תדירה באותם אזורים.

4. משך הנסיעה כתלות בשעת התחלתה

נסיעות המתחילות בשעות הקרובות לשעות העומס, כמו 7:00 בבוקר ו-16:00 אחר הצהריים, נמשכות זמן רב יותר בממוצע. בשעות אלו, תנועת הרכבים על הכבישים היא אינטנסיבית יותר, מה שמוביל לעיכובים ולהתארכות משך הנסיעה.

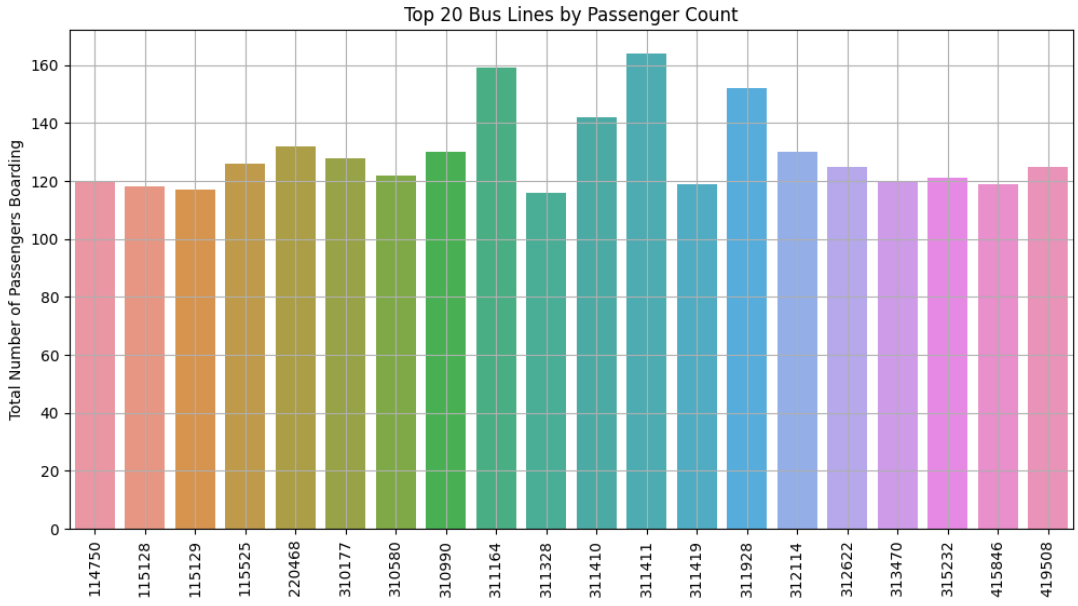
מנגד, בשעות הבוקר המוקדמות (לפני 7:00) ובשעות הערב המאוחרות (אחרי 19:00), משך הנסיעות קצר יותר באופן ניכר. הפיזור הרחב של משך הנסיעה בשעות העומס מעיד גם על שונות גדולה יותר ואי-ודאות לגבי משך הנסיעה, מה שמצביע על פוטנציאל לעיכובים לא צפויים.

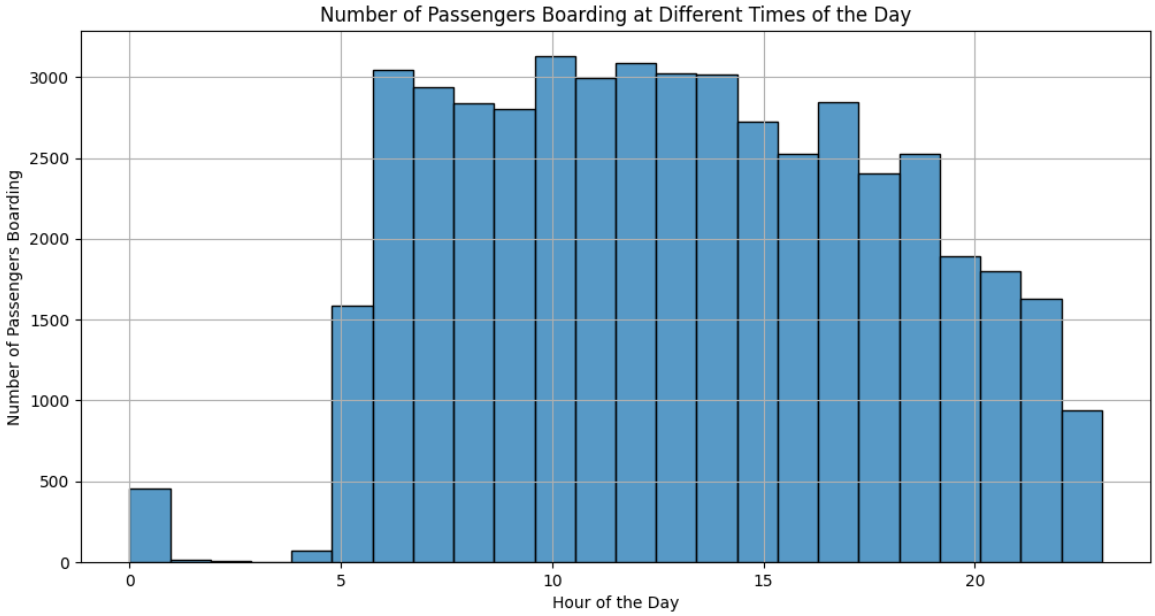
בשעות העומס יש יותר נקודות חריגות המייצגות נסיעות ארוכות במיוחד, לעומת השעות השקטות יותר בהן התנועה זורמת וחלקה.

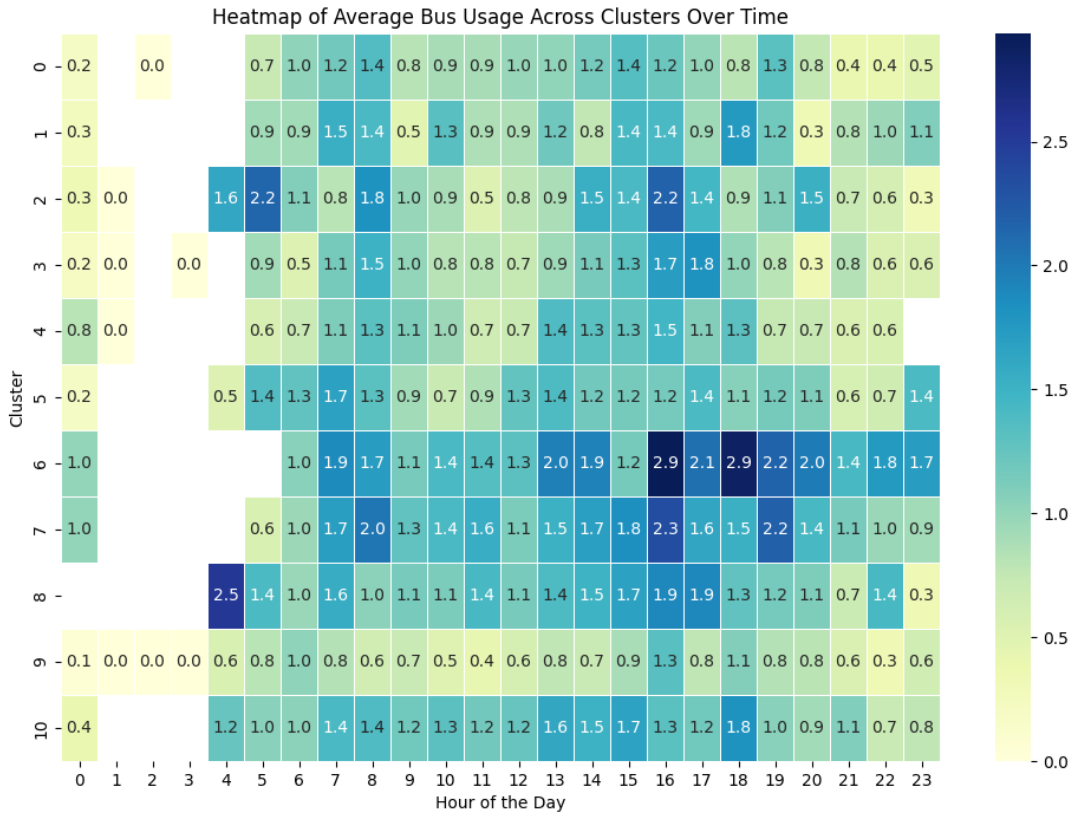
<u>המלצות</u>: בשעות העומס, מומלץ לשקול הגברת תדירות האוטובוסים על מנת להפחית את העיכובים ולהקטין את משך הנסיעה לנוסעים. כמו כן, מומלץ להתמקד בניהול עיכובים לא צפויים באמצעות טכנולוגיות ניטור וניהול תנועה.

תרשימים

Task 3.1 - Predicting Passenger Boardings at Bus Stops






Heatmap of Average Bus Usage Across Clusters Over Time
Cluster
Hour of the Day
0 0.2 0.0 0.7 1.0 1.2 1.4 0.8 0.9 0.9 1.0 1.0 1.2 1.4 1.2 1.0 0.8 1.3 0.8 0.4 0.4 0.5
1 0.3 0.9 0.9 1.5 1.4 0.5 1.3 0.9 0.9 1.2 0.8 1.4 1.4 0.9 1.8 1.2 0.3 0.8 1.0 1.1
2 0.3 0.0 1.6 2.2 1.1 0.8 1.8 1.0 0.9 0.5 0.8 0.9 1.5 1.4 2.2 1.4 0.9 1.1 1.5 0.7 0.6 0.3
3 0.2 0.0 0.0 0.9 0.5 1.1 1.5 1.0 0.8 0.8 0.7 0.9 1.1 1.3 1.7 1.8 1.0 0.8 0.3 0.8 0.6 0.6
4 0.8 0.0 0.6 0.7 1.1 1.3 1.1 1.0 0.7 0.7 1.4 1.3 1.3 1.5 1.1 1.3 0.7 0.7 0.6 0.6
5 0.2 0.5 1.4 1.3 1.7 1.3 0.9 0.7 0.9 1.3 1.4 1.2 1.2 1.2 1.4 1.1 1.2 1.1 0.6 0.7 1.4
6 1.0 1.0 1.9 1.7 1.1 1.4 1.4 1.3 2.0 1.9 1.2 2.9 2.1 2.9 2.2 2.0 1.4 1.8 1.7
7 1.0 0.6 1.0 1.7 2.0 1.3 1.4 1.6 1.1 1.5 1.7 1.8 2.3 1.6 1.5 2.2 1.4 1.1 1.0 0.9
8 2.5 1.4 1.0 1.6 1.0 1.1 1.1 1.4 1.1 1.4 1.5 1.7 1.9 1.9 1.3 1.2 1.1 0.7 1.4 0.3
9 0.1 0.0 0.0 0.0 0.6 0.8 1.0 0.8 0.6 0.7 0.5 0.4 0.6 0.8 0.7 0.9 1.3 0.8 1.1 0.8 0.8 0.6 0.3 0.6
10 0.4 1.2 1.0 1.0 1.4 1.4 1.2 1.3 1.2 1.2 1.6 1.5 1.7 1.3 1.2 1.8 1.0 0.9 1.1 0.7 0.8
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
0.0 0.5 1.0 1.5 2.0 2.5

Task 3.2 - Predicting Trip Duration