**NY Rental Properties Pricing**

בחרנו בדאתאסט מאוד מעניין המדבר על מחירי הדירות להשכרה בניו יורק, הדאטאסט מציג תכונות מעניינות לגבי הדירות בניו יורק כמו המחיר, קו אורך וקו רוחב וכו', אחרי בדיקה והבנה מעמיקה על מה מדבר הדאטאסט החלטנו לחזות את סוגי הדירות בעזרת הנתונים שיש לנו.

הרצנו 8 איטרציות בכל איטרציה השתמשנו בקומבינציות שונות של עיבוד מקדים שלמדנו לאורך הסמסטר ואחרי השוואה בין תוצאות האיטרציות החלטנו איזו שיטות היו עדיפות ונותנות לנו את התוצאה הכי טובה למה שאנחנו מחפשים.

**מצאנו בסופו של דבר שהשיטות הכי טובות לנו בעיבוד מקדים היו (שיטת העיבוד – שם העמודה):**

Bin means – days\_occupied\_in\_2019,

Decimal Score- reviews\_per\_month + avaliability\_2020,

Z-score – price + minimum\_nights,

MinMax - number\_of\_reviews

**והורדנו את העמודות הלא רלוונטיות:**

id, F1, latitude, longitude, neighbourhood

Id – index column

F1 – some identification column, not consistent and non-informative

Latitude-longitude : location of the asset

Neighborhood – name of the neighborhood that the asset is in

כל העמודות הללו לא רלוונטיות ולא יכולות לתרום לנו למידע.

**מסקנות לגבי העיבוד המקדים:**

אנחנו חושבים שהעיבוד המקדים היה שלב קריטי וחשוב ובלי זה לא יכלנו לחזות דברים ולנתח דברים בצורה נכונה, העיבוד המקדים עזר לנו לסדר את הנתונים על מנת להבין מה אנחנו צריכים ומה אנחנו לא צריכים מהעמודות שיש לנו עשינו שיטות שלמדנו במהלך הסמסטר וזה היה שלב חשוב מאוד להתקדמות בפרויקט.

**המודל הכי טוב שהיה לנו:**

- Random Forest

המודל נתן לנו רמת דיוק של 0.8277 וזה ביחד עם שיטות העיבוד המקדים לעיל, זאת היתה התוצאה הכי גבוהה שקיבלנו, בהשוואה ל (majority rule = 0.53395) שלנו רמת הדיוק שקיבלנו היתה יותר טובה, לכן אפשר להסיק שהמודל נתן לנו תוצאות יותר טובות מאשר ללכת עם ה-(majority rule).

**מסקנות לגבי המודל:**

המודל הזה מורכב ממספר עצים המורכבים מהמצביעים השונים בקבוצת האימון, והיה לו יכולת מעולה לזהות ולחזות את התגובה הנכונה של הנתונים בצורה מדויקת. כמו כן, המודל היה יכול לזהות את המאפיינים החשובים ביותר לקביעת התוצאה.

**מגישים : לידור פלדמן 209297035, מוחמד אבו ג'עפר 213169915**