סיכום

סיווג ביקוש להשכרת אופניים

מנחה: סיוואר

מנטור: אין

אנשי הצוות: רואי שם טוב ואייל גרינפלד

מבוא ①

הסבר על הדומיין

חיזוי הביקוש להשכרת אופניים מושפע ממגוון גורמים, כמו: מזג אוויר, זמן ומקום. לשם החיזוי אספנו נתונים אודות השכרת אופניים, נתונים שנאספו ע"י חיישנים המותקנים בתחנות ההשכרה. מטרת החיזוי היא לאפשר ניהול אופטימלי של מלאי האופניים, כך שלא יווצרו עומסים בתחנות מסויימות.

מידע כללי על מאגר הנתונים

מאגר הנתונים נאסף בחודשים אוקטובר, נובמבר ודצמבר 2023 ע"י חברת Lyft, המפעילה את מערכת Citi Bike, ואחראית על איסוף הנתונים.
בסוף תהליך ה- EDA בסיס הנתונים כולל כ- 4,000,000 נתונים ו- 8 פיצ'רים.

שאלת המחקר והבעיה העסקית

שאלת המחקר: מהם הגורמים המשפיעים על הביקוש להשכרת אופניים?

הבעיה העסקית: כיצד ניתן לייעל את ניהול מלאי האופניים, כך שלא ייווצרו עומסים ומחסור באופניים בתחנות מסוימות, בעוד שאחרות יהיו מלאות?

התרומה לענף ולעולם: יצירת מודל מוצלח שיחזה את הביקושים בתחנות השונות יסייע למפעילים לנייד אופניים מתחנות עם ביקוש נמוך לגבוה. כך יותר אנשים יבחרו להתנייד באופניים והזיהום יקטן.

סיכום שלב 2
FE+EDA

טבלת הפיצ'רים

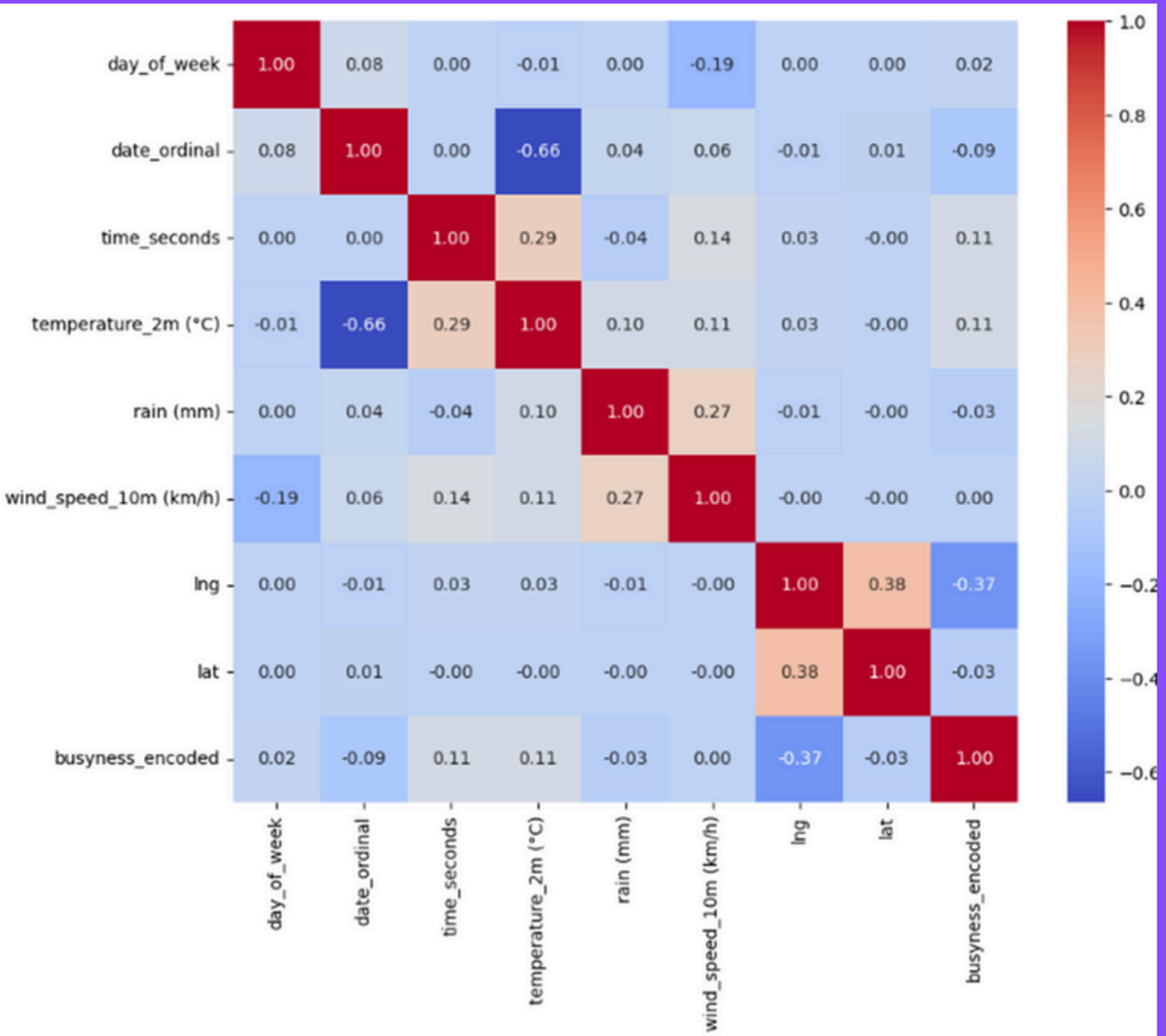
פיצ'ר	משמעות	כמות נתונים לאחר EDA מתוך כמות מקורית
Temperature	הטמפרטורה בניו יורק במעלות צלזיוס (בשעה ובתאריך המתאים)	2,032,979/4,220,402
Rain	כמות הגשם במ"מ שירדה בשעה מסוימת בניו יורק	2,032,979/4,220,402
Wind Speed	מהירות הרוח בקמ"ש	2,032,979/4,220,402

טבלת הפיצ'רים

כמות נתונים לאחר EDA מתוך כמות מקורית	משמעות	פיצ'ר
2,032,979/4,220,402	קווי אורך/רוחב של מיקום התחנה	Latitude/Longitude
2,032,979/4,220,402	השעה (העגולה)	Time
2,032,979/4,220,402	היום בשבוע (ראשון, שני...)	Day of Week
2,032,979/4,220,402	תאריך (ללא שנה)	Date

גרף קורלציות

corr(x,y)



- התאריך משפיע על הטמפרטורה.
- מידת העומס מושפעת מ-lng.
- מידת העומס מושפעת במידה מועטה מאוד משאר הפיצ'רים.

מסקנות עיקריות משלב ה- EDA + FE

- הקורלציות בין הפיצ’רים נמוכות מאוד (גם עם עמודת המטרה), ולכן ייתכן שהקשרים אינם לינאריים.
- התפלגות עמודת המטרה אינה אחידה.

3 מדדי דיוק

Accuracy

- מתאר את אחוז הסיווגים הנכונים.
- עשוי להטעות במקרים של חוסר איזון בנתונים, ולכן שימושי במקרים שבהם יש איזון בין הקטגוריות, מה שלא מתקיים אצלנו.

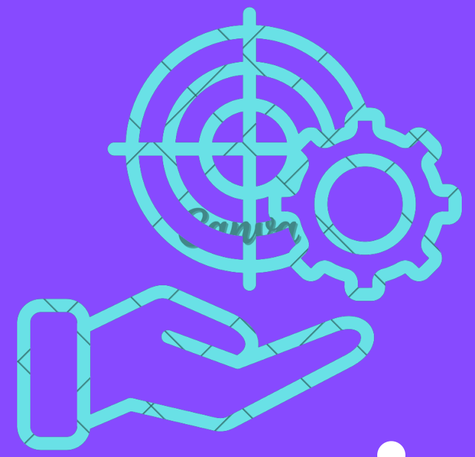
Precision

- מתאר את אחוז הסיווגים הנכונים מתוך כל הסיווגים של המודל לקטגוריה מסוימת.
- חשוב כאשר העלות של טעות מסוג False Positive גבוהה.
- במקרה שלנו, בודק אם המודל מנבא "עומס גבוה" כאשר בפועל אין עומס.
- לכן, המדד מתאים לנו כדי למנוע זיהוי יתר של עומס בתחנה, כך שלא נקצה משאבים רבים לתחנה שלא לצורך.

Recall

- מתאר את אחוז הסיווגים הנכונים מתוך כלל המקרים האמיתיים של הקטגוריה.
- מדד קריטי במקרים שבהם העלות של False Negative גבוהה.
- במקרה שלנו, בודק אם המודל מנבא שאין עומס בתחנה בזמן שבפועל יש עומס.
- לכן, המדד מתאים לנו כדי לזהות את כל מקרי העומס, כך שנוכל למנוע אותם מראש.

F1-Score



- מהווה ממוצע הרמוני של Precision ו- Recall.
- במקרה שלנו, משמש כדי להבטיח שהמודל לא יפספס עומס אמיתי (Recall), אך גם לא ינבא עומס מיותר (Precision).
- לכן, זהו המדד העיקרי בו השתמשנו להערכת ביצועי המודלים.
- אנחנו השתמשנו ב- weighted f1-score בשל חוסר האיזון בעמודת המטרה ומשום שמדובר בסיווג רב קטגורי.

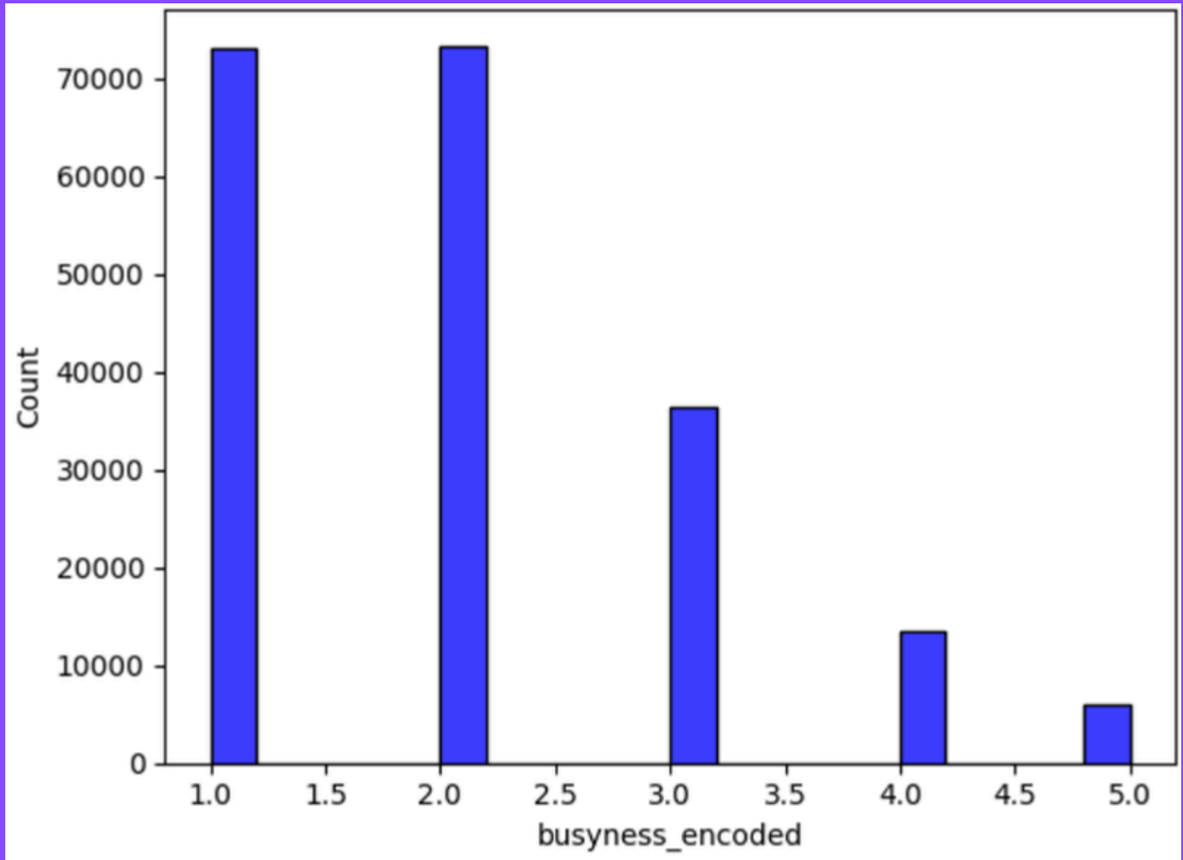
4 נתוני בדיקות

Train-Test Split

- תחילה הפחתנו את מספר הרשומות ל- 400,000 במקום 4,000,000. לשם כך נעזרנו ב- Stratified Sampling ששומר על התפלגות עמודת המטרה.
- לאחר מכן חילקנו את הנתונים כך ש-80% ישמשו לאימון ו-20% למבחן.
- בנוסף, חילקנו את סט האימון ל-5 folds באמצעות Cross Validation, כך שכל שכל fold שימש פעם אחת כסט בדיקה.

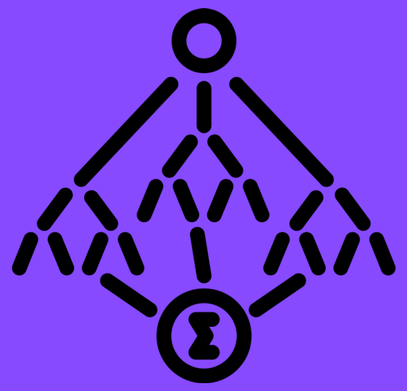
Test Set

- הטסט סט עליו ביצענו את ההערכה וההשוואה בין המודלים כלל 400,000 רשומות.



5 סוגי מודלים

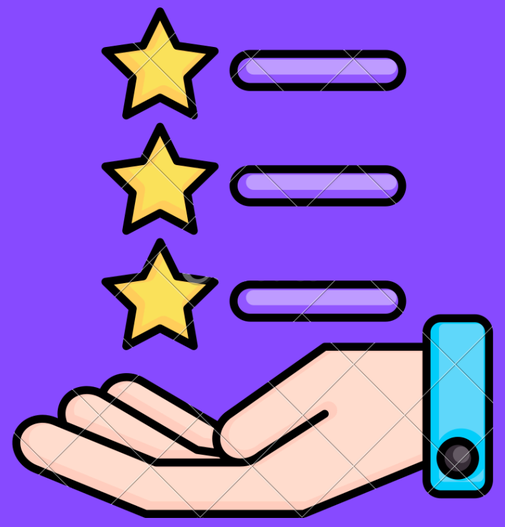
Random Forest



- מודל היוצר מספר עצי החלטה, כאשר כל עץ נבנה על מדגם אקראי מהנתונים המקוריים (Bootstrap Sampling).
- בכל עץ, האלגוריתם בוחר תכונות אקראיות מתוך התכונות הקיימות כדי לחלק את הנתונים.
- במקרה של סיווג (Classification): כל עץ "מצביע" על קטגוריה, והקטגוריה בעלת מספר ההצבעות הגבוה ביותר נבחרת (Voting).

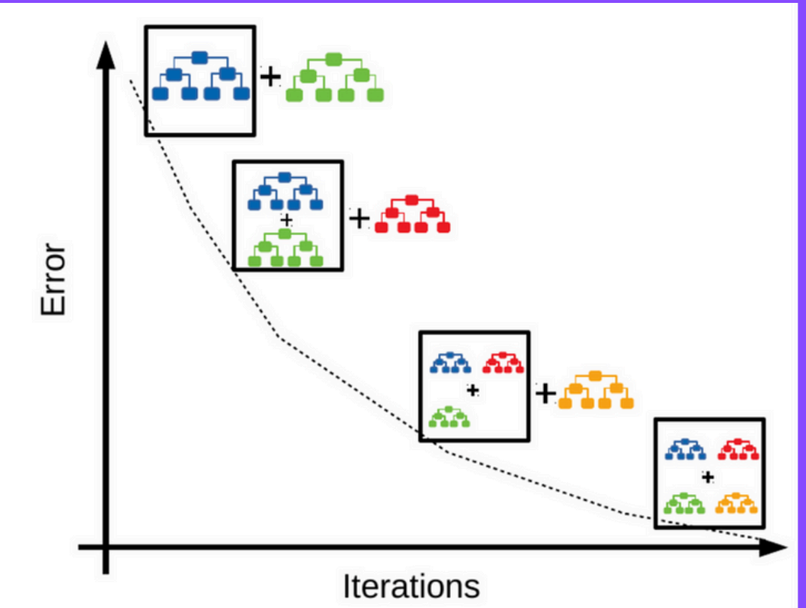
יתרונות - Random Forest

- **חוסר איזון:** המודל מתאים למקרים של נתונים לא מאוזנים.
- **מניעת Overfitting:** בזכות שילוב של עצים רבים, האלגוריתם נוטה להיות פחות רגיש ל-Overfitting בהשוואה לעץ החלטה בודד.
- **נתונים מורכבים:** המודל מתאים לתרחישים עם נתונים מרובי תכונות ונתונים שאינם ליניאריים.



Gradient Boosting

- **מודל ראשוני:** האלגוריתם מתחיל על ידי התאמה של עץ החלטה קטן לנתונים.
- **חישוב השגיאה של המודל.**
- **בניית מודל חדש:** בניית עץ החלטה נוסף כך שלא יחזור על השגיאות שהתקבלו בשלב הקודם.
- **חזרה על התהליך.**

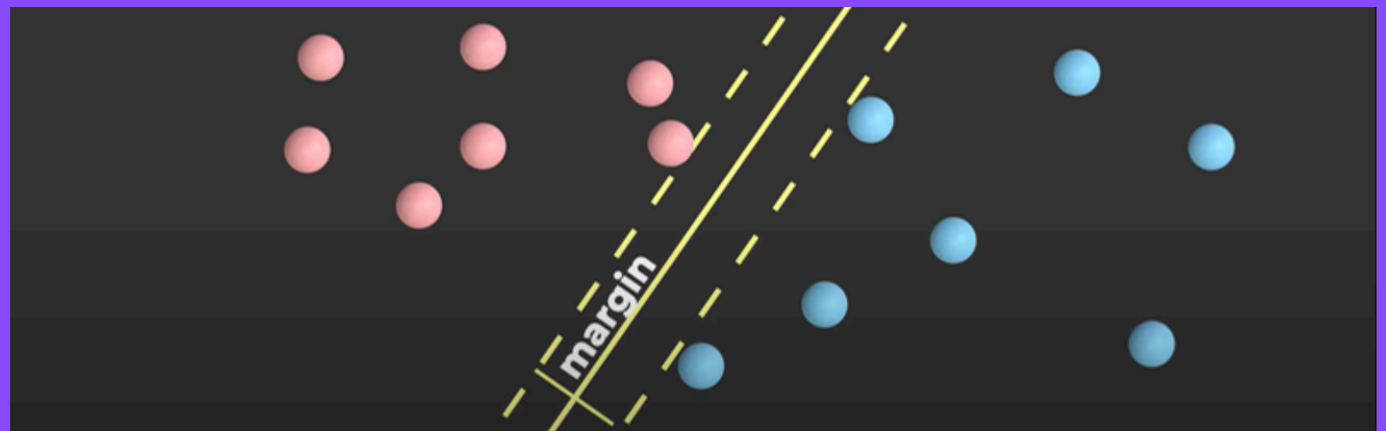


יתרונות - Gradient Boosting

- **טיפול בנתונים מורכבים:** Gradient Boosting מפיק תוצאות מדויקות מאוד גם במקרים שבהם הנתונים מורכבים או לא לינאריים.
- **טיפול בנתונים לא מאוזנים:** האלגוריתם מתאים היטב לבעיות שבהן יש קטגוריות לא מאוזנות.

SVM

- המטרה של SVM היא למצוא היפר-מישור (Hyperplane, משטח) שמפריד בין הקטגוריות בצורה מיטבית.
- טיב ההפרדה נקבע לפי המרווח (margin) בין המשטח לנקודות הקרובות ביותר. המודל מחפש את ההיפר-מישור עם המרווח המקסימלי כדי למנוע overfitting.
- מתאים גם לנתונים שאינם ניתנים להפרדה לינארית (היתרון).



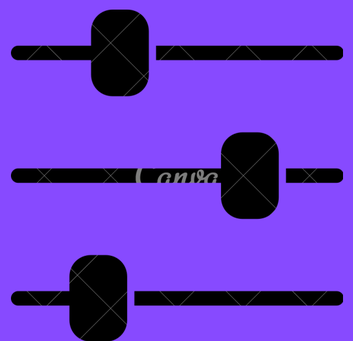
שימוש ב- Voting

- השתמשנו ב- voting בין המודלים Random Forest ו- XGBClassifier.
- עשינו שימוש ב- soft voting, כך שהסיווג התקבל לפי המחלקה שהמודלים היו הכי בטוחים בה.
- השתמשנו בשיטה זו ממספר סיבות:

1. **<u>מניעת אובר-פיטינג:</u>** Random Forest יעיל יותר בכך לעומת XGBClassifier, כך שהשילוב ביניהם מסייע למנוע זאת.
2. **<u>מניעת טעויות:</u>** כל מודל יכול לתקן את השני.
3. **<u>זיהוי דפוסים מגוונים:</u>** XGBoost יכול להתמודד טוב עם נתונים לא לינאריים, בעוד ש- Random Forest יוכל לזהות דפוסים פשוטים יותר.

ערכי hyper-parameters

- **max_depth=None** - אין הגבלה על עומק העץ (מאפשר לעץ ללמוד את כל הדפוסים המורכבים בנתונים).
- **max_features='log2'** - מספר הפיצ'רים המקסימלי שמהם כל עץ יכול לבחור בחלוקה (מקטין Overfitting).
- **min_samples_leaf=4** - המספר המינימלי של דגימות הנדרשות כדי להפוך לצומת עלה (מקטין Overfitting).
- **min_samples_split=10** - המספר המינימלי של דגימות הנדרשות בצומת כדי לבצע חלוקה (מקטין Overfitting).
- **n_estimators=180** - מספר העצים ביער.



6 ניסיונות לשיפור

ניסיונות לשיפור המודל

הניסיון	התוצאה	מדוע זו התוצאה?
הורדת הפיצ'ר wind_speed	שיפור של 5% במדד f1-score (47%)	משום שהקורלציה בין פיצ'ר זה לעמודת המטרה נמוכה מאוד
כיוונון הייפר-פרמטרים (העלאת n_estimators)	שיפור של 2% במדד f1-score (49%)	משום שיותר עצים ביער יכולים לשפר את הדיוק
Oversampling	ללא שינוי במדד f1-score	משום שלמודל היו מספיק נתונים לאימון מכל מחלקה

ניסיונות לשיפור המודל - המשך

הניסיון	התוצאה	מדוע זו התוצאה?
voting בין Random Forest ל- XGBClassifier	שיפור של 2% במדד f1-score (51%)	מניעת overfitting, מניעת טעויות וזיהוי מגוון דפוסים

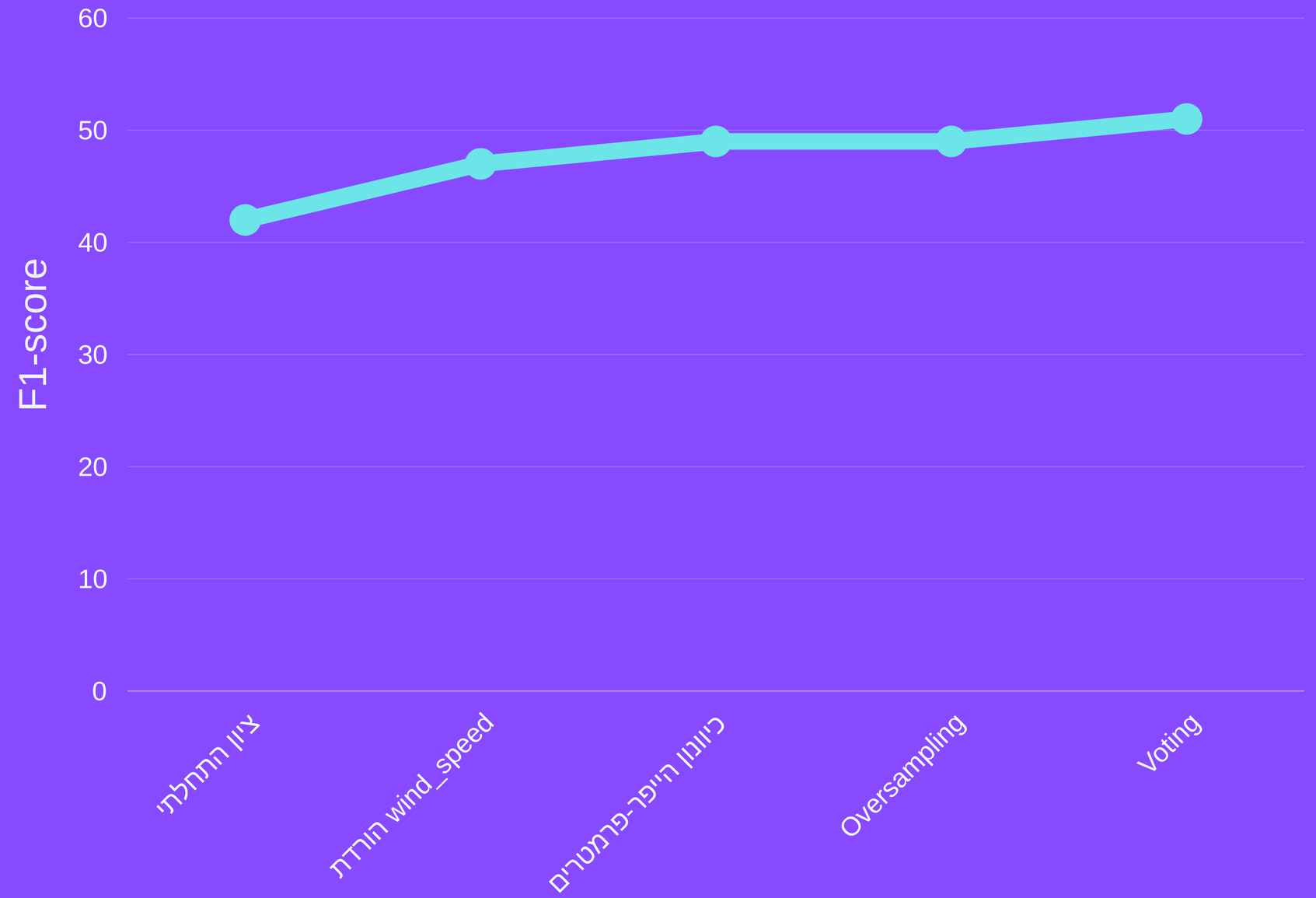
ניסיונות לשיפור המודל - תובנות

- מה מאפיין את כל השיפורים? שיפור איכות הנתונים (Oversampling והסרת wind_speed) ושיפור היכולות של המודל (Voting וכיוונון הייפר-פרמטרים).
- מדוע לא הסרנו את wind_speed בשלב ה- EDA+FE? משום שרק בשלב זה יכולנו לדעת שתרומתו שולית, כנראה כי המודל כבר למד מידע דומה מפיצ'רים כמו rain או temperature.
- אין תאימות בין הפיצ’רים של המודל לבין הקורלציות.

7

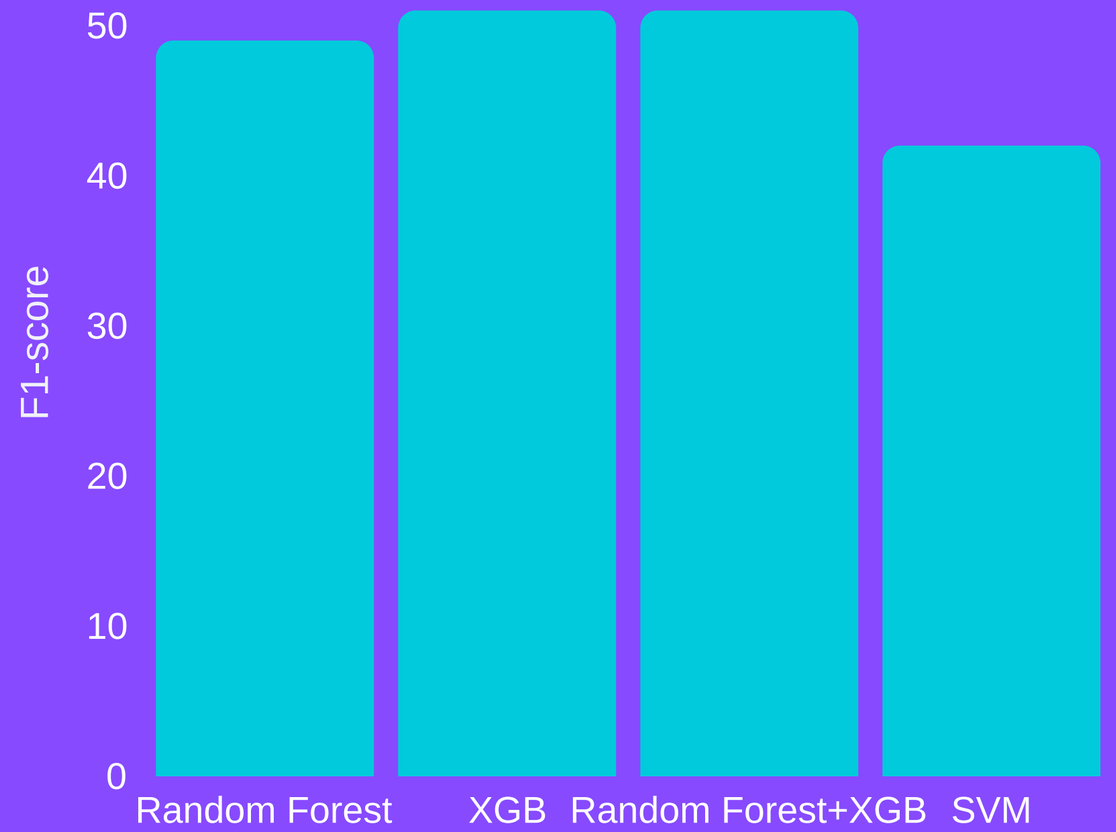
סיכום
הערכת
מודלים

ניסיונות לשיפור המודל - גרף



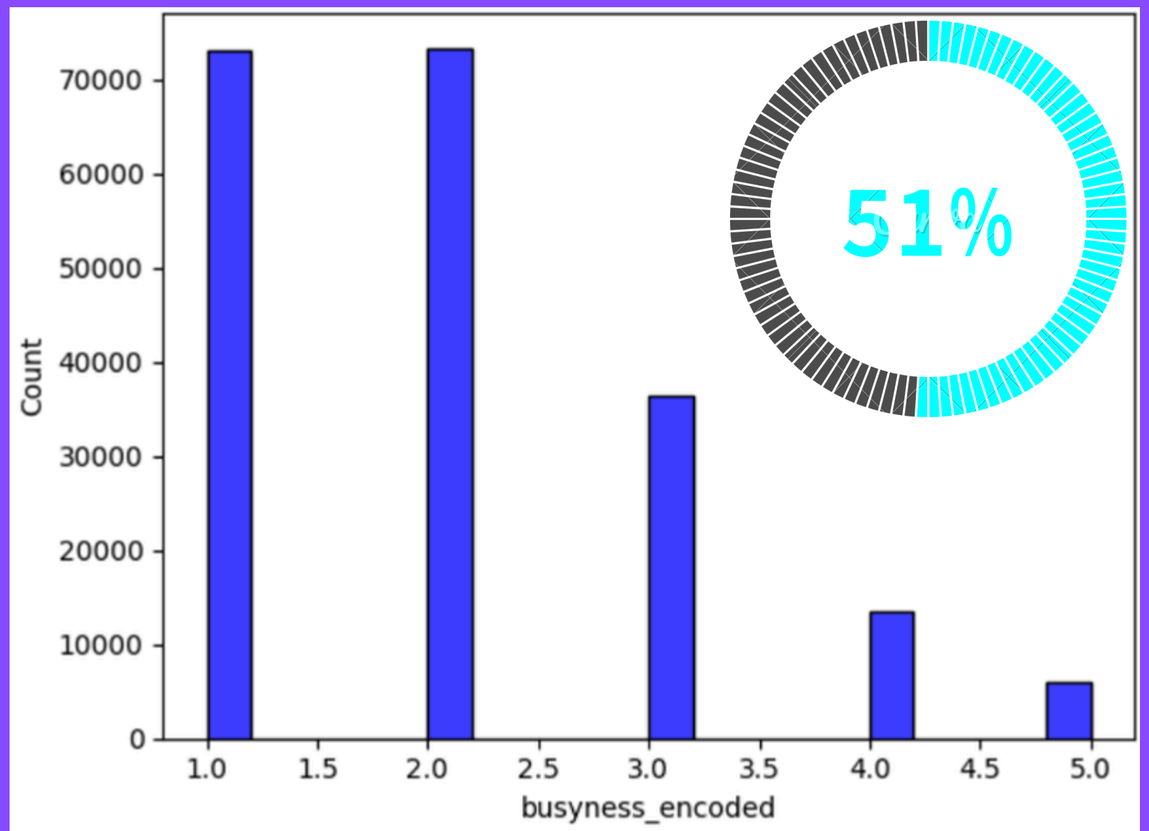
השוואה בין מודלים

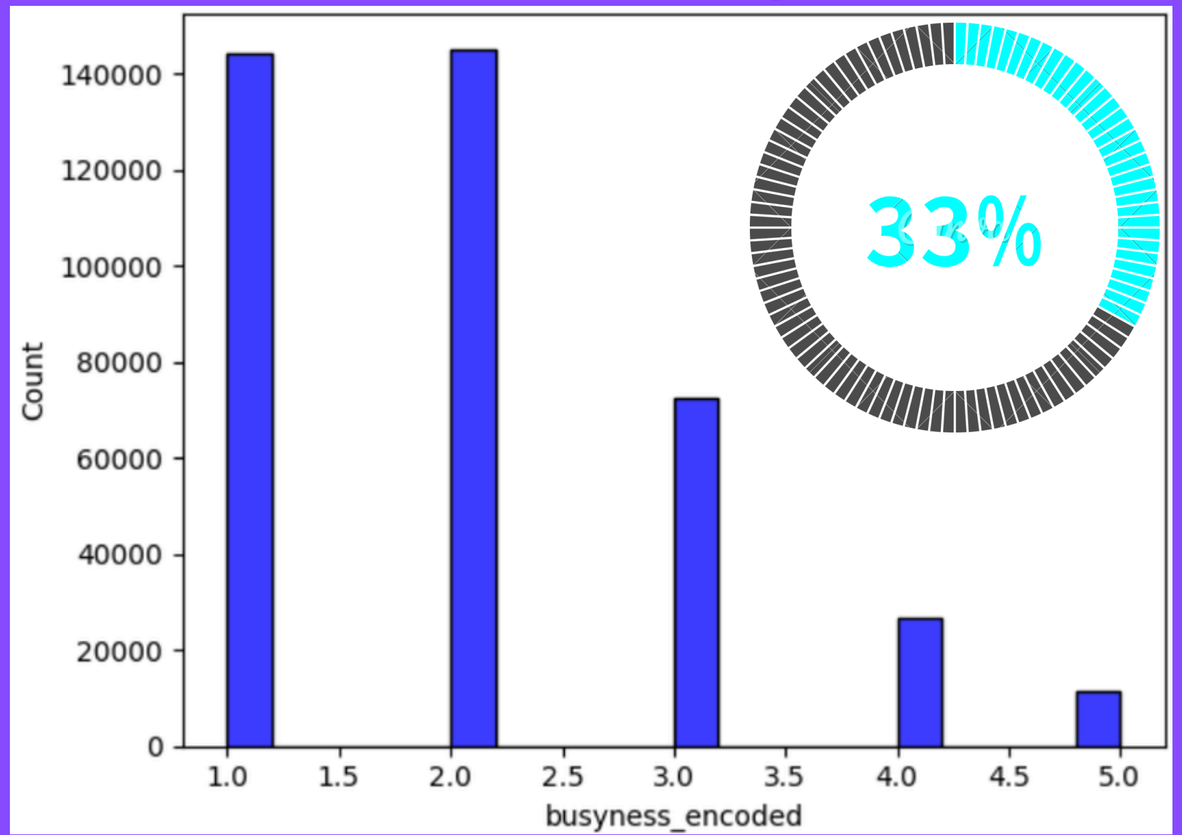
- XGBClassifier טוב בגלל היכולת שלו ללמוד קשרים מורכבים.
- Ensemble Voting טוב בגלל שהוא משתמש בחוזקות של שני המודלים יחד.
- SVM פחות טוב כנראה כי דרוש כיוונון של ההיפר-פרמטרים שלו.



השוואה בין טסט סטים

- ייתכן שההבדלים בין הטסט סטים נובעים מערכי פיצ'רים שלא נראו בסט האימון, כמו מהירות רוח גבוהה במיוחד או טמפרטורות מחוץ לטווח שסט האימון כלל.





8 סיכום הפרויקט

נושאים חדשים שלמדנו

- שילוב של מסדי נתונים שונים שלקחנו מהאינטרנט אל קובץ מידע אחד ומאורגן.
- שימוש והבנה של SVM, Gradient Boosting, Voting.

לקחים מהעבודה

הבנו שהרבה יותר אפקטיבי וחסכוני בזמן לחלק את העבודה בינינו, כך שגם כאשר אחד מאיתנו לא פנוי לעבוד השני יוכל להתקדם ולהמשיך לעבוד על הפרויקט (מה שלא מתאפשר בעבודה ביחד במקביל).

דברים לשיפור

- ניתן לקיים תחרות בין קבוצות על מדדי דיוק טובים יותר.
- ניתן לפצל את יום הצגת המצגות לשניים, כך שלא נצטרך לצפות במצגות במשך 4 שעות.

דברים לשימור

- **חופש הפעולה:** היכולת שלנו לבחור בהתאם לרצוננו, משלב איסוף הנתונים ועד בניית המודל, שיפרה מאוד את עצמאותנו בתחום.
- **הליווי האישי:** ההכוונה והתמיכה שקיבלנו מסיוואר עזרו לנו מאוד להתקדם בפרויקט.

כמה נהנינו?

נהנינו מאוד מביצוע הפרויקט, מכיוון שלמדנו דברים חדשים והתנסינו בבניית פרויקט מההתחלה ועד סופו ללא הוראות ברורות.

עם זאת, חווינו גם אתגרים רבים שהיו פחות מהנים, אך גם מהם למדנו המון, והפרויקט בכללותו היה מלמד ומהנה.

הקשיים העיקריים

- **ניהול הזמן** - ביצוע הפרויקט היה במהלך תקופת מבחנים ועבודות בבית הספר, ולכן היה מאתגר למצוא זמן להתקדם בפרויקט.
- **נושאים חדשים** - היה מאתגר ללמוד נושאים חדשים לגמרי לבד, כמו SVM ו- Voting, אך התגברנו על כך בעזרת סרטונים מהאינטרנט.

ההבדל בין הפרויקט לאבני הדרך

בניגוד לאבני הדרך בכיתות י' ו- י"א, כחלק מהפרויקט יכולנו לבחור נושא שמעניין אותנו. בנוסף, בפרויקט זה מצאנו את מסדי הנתונים בעצמנו ולאורך כל העבודה התקדמנו באופן עצמאי וללא הוראות מפורטות. כמו כן, הפרויקט כלל גם הצגה וליווי של מנחה.

תודות

בהזמנות זו נרצה לומר תודה רבה לסיוואר המנחה שלנו שמדי שבוע דאגה להתעדכן בהתקדמותנו ותמיד הייתה מוכנה לעזור. בנוסף, רצינו להגיד תודה לכל צוות מגשימים AI (ויקי, תמר, בן, עינב, שרה...) שאפשרו לנו לקבל את ההזדמנות המדהימה הזו וללמוד את הנושא של המחר (וההווה).

תודה על ההקשבה