**פרויקט האקתון - 2021**

**משימה - חיזוי רווחים וציון סרטים**

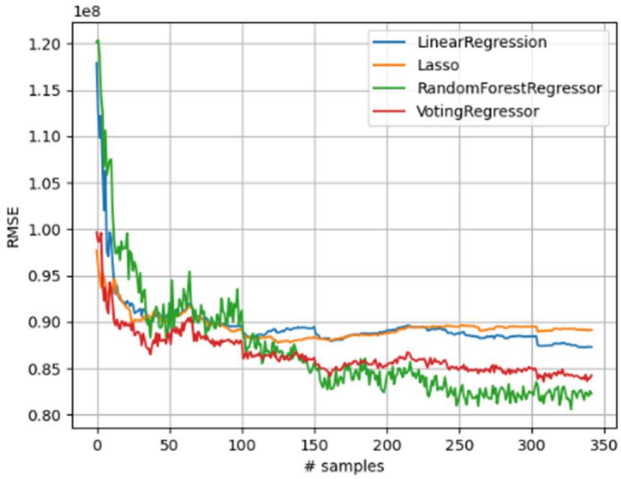
**מגישים: ניר גבריאלוב, הדר שרביט, יפעת חדד, שלום קצ'קו**

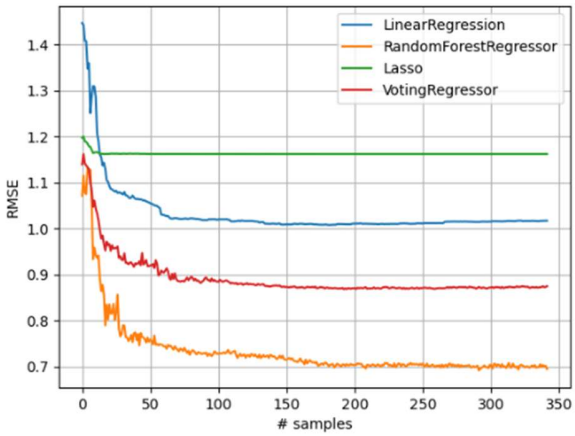
- **תיאור הדאטה**: סט הנתונים שלנו היה מורכב מכ - 6000 דגימות וכ - 20 פיצ'רים המתארים כל דגימה כאשר חלקים משתנים קטגוריים. אחד האתגרים העיקריים בנתונים שקיבלנו היו המשתנים הקטגוריים אשר לא מאופניינים ביחס סדר מובהק בינהם. החלטנו להשתמש ב - *one hot encoding* ע"מ לאפיין את הפיצ'רים בצורה מספרית. קושי נסף היה ההתמודדות עם הפארסינג של קבצי ה - *json*, חלקם הכילו מידע שדרש עיבוד מקדים משמעותי לפני שהיה ניתן להגדיר בעזרתו פיצ'ר בעל משמעות.

- **עיבוד וניקוי הנתונים**: התחלנו משלבי ניקוי ראשונים והסרנו ערכי זבל לדוגמת תאים עם Nan או ערכים שליליים עבור משתנים שאינם יכולים להיות שליליים במציאות. בחרנו להסיר שורות שיש בהן 4 Nan או יותר מכיוון ששורות אלו לא הביאו לנו מידע מספיק. בעמודות שהכילו את הערך 0 ולזה לא אפשרי במציאות (לדוגמת זמן ריצה) החלפנו את האפס בחציון של העמודה בכדי להמנע מהסטה של הנתונים. בתהליך ההמרה של משתנים קטגורים חילצנו את חברת ההפקה, ז'אנר והשפה דגמנו את המשתנים החוזרים המשפיעים ביותר בכל תחום ויצרנו וקטור one hot, אחרת סיווגנו כ- other. לאחר מכן התמודדנו עם הטקסט החופשי, ביצענו סקירה של כל המילים הקיימות (להוציא מילות stopwords) ואז ביצענו stemming בשביל לקבל את שורשי המילים. קיבלנו כ - 18000 משתנים חדשים. ביצענו היסטוגרמה ובחרנו סף כלשהו בכדי לסנן חלק מהמשתנים. לאחר ההרצה ראינו שהשימוש במילים לא משנה בצורה משמעותית את תוצאות ולכן ויתרנו על המילים לטובת מימד קטן יותר. מהתאריכים הנתונים יצרנו שני מתשנים קטגוריים, היום בשבוע שבו הסרט יצא והאם תאריך היציאה בשבוע שלפני חג.

- **בחירת מודל למידה**: תחילה בחרנו מספר מודלים פשוטים, הקצנו חלק קטן מהנתונים שלנו לטובת בניית המודלים הפשוטים ושרטוט ויזואלי של הנתונים והשגיאה. לאחר מכן השוונו בין המודלים הקיימים למודלים עם רמת מורכבות גבוהה יותר ומתוכם בחרנו את המודלים בעלי הביצועים הטובים ביותר. עבור המודלים שבחרנו הגדרנו כלל ועדה בכדי להקטין את השגיאה בתוחלת. בשלב בו קיבענו את מודל הלמידה ניסינו לבצע כונון של ההייפר-פרמטרים (עומק העץ וכמות המדידות המינימלית בחלוקה) ע"י שימוש ב - kfold. המודלים אשר השתמשנו בהם: Lasso, Ridge, LinearRegression, RandomForest, RegressionTree כאשר בכלל ההחלטה שלנו השתמשנו רק ברגרסיה ליניארית, RandomForest ולאסו. לאחר בחינת הממצעים החלטנו להשתמש ב - RandomForestRegressor.

- **חיזוי שגיאה**: ביצענו אימון של המודל על 60% מהנתונים שקיבלנו, לאחר מכן הרצנו את המודל המאומן על 20% מהנתונים שהם ה - validation שלנו והשתמשנו בסט זה בכדי לקבוע היפרפרמטרים. אחרי קבעת ההיפר פרטמרים הרצנו על 20% הנותרים של הנתונים בתור test כאשר ציפינו שהשגיאה תהיה דומה לזאת של ה - validation. מצורף גרף השגיאה של כל אחד מהמודלים, משמאל עבור הרווחים ומימין עבור הציון:


1e8
1.20
1.15
1.10
1.05
1.00
0.95
0.90
0.85
0.80
RMSE
0
50
100
150
200
250
300
350
# samples
LinearRegression
Lasso
RandomForestRegressor
VotingRegressor


1.4
1.3
1.2
1.1
1.0
0.9
0.8
0.7
RMSE
0
50
100
150
200
250
300
350
# samples
LinearRegression
RandomForestRegressor
Lasso
VotingRegressor