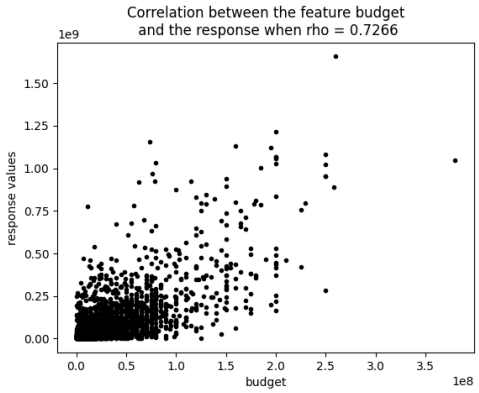
האקתון IML 2021
עדי, אורן, הילה וגבריאלה
משימה 1 - Movies

★ Describe the dataset, and any challenging characteristics it has and describe (briefly) the data cleaning and preprocessing?

הדאטה מורכב מכ-7000 סרטים כאשר כל סרט מכיל 22 פיצ'רים. עברנו על כל הפיצ'רים וניסנו להסיק מה כל נתון יכול לתרום לפרדיקציה שלנו.
ראשית, יש מספר פיצ'רים בעלי יחס סדר מוכר (לדוגמא: תקציב, מספר הצבעות). בנוסף מצאנו מספר פיצ'רים שהשפעתם על הדאטה היתה בינארית, כלומר קיום או לא. (לדוגמא: האם ישנו לינק ל homepage, האם הסרט משתייך לקולקשיין כלשהו ועוד).
בחרנו להמיר את התאריך לשני פיצרים שונים כאשר אחד מסמל את החודש בו הסרט יצא לאור והשני מייצג את מספר הימים שעברו מאז שהסרט יצא ועד היום.
לאחר מכן, המרנו מספר עמודות ל-Dummy-Variables (לדוגמא: ז'אנר, שחקנים וכו) אך נתקלנו בבעיה של מספר רב מידי של Dummies ולכן יצרנו פונקציה אשר תחשב את ה-Top-Dummy-Values ורק אותם הכללנו בתור פיצרים, שהרי הם בעלי החשיבות המרכזית.
בנוסף, היינו צריכים לטפל במקרים בו היו חסרים ערכים של הפיצרים, ניסינו למצוא האם יש ערכי outliers קיצוניים ובמקרים כאלו החלפנו ערכים חסרים בחציון.

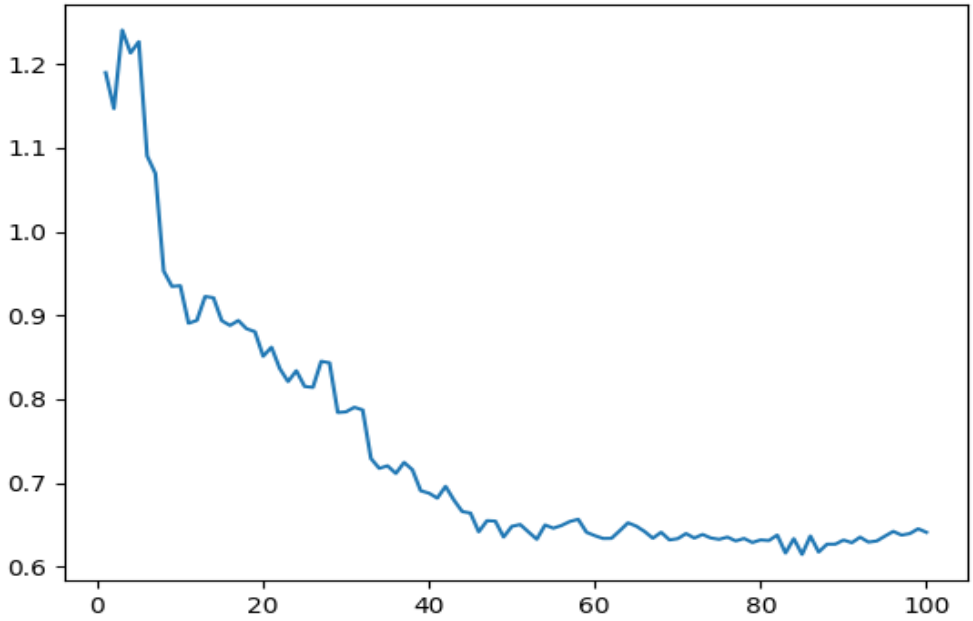


יצרנו גרפים המתארים את הקורלציה בין פיצ'רים מסוימים ל response וזה אפשר לנו להסיק על הרלוונטיות של הפיצ'רים המסויימים לבעיה שלנו. כפי שניתן לראות בגרף, budget כנראה הוא פיצ'ר רלוונטי עבור חיזוי הrevenue.

★ Describe the considerations that guided your design of learning systems?

מדובר בבעיה רגרסיה ולכן חשבנו על מודלים שמטפלים בסוג זה של מערכות למידה - רגרסיה לינארית (עם רגולריזצית ridge ו-lasso), יער החלטה של רגרסיה (כולל רגולריזציה על גודל העץ), והשווינו מודלים אלה גם למערכות מוכנות מראש של ספריית scikit-learn כמו GradientBoostingRegressor ו- AdaBoostRegressor. כמו כן, לכל אחת מהבעיות (אחת - מציאת הרווח הצפוי, והשנייה - מציאת הדירוג הצפוי), התאמנו בחירת פיצ'רים שלה- תהליך שהחל באופן ידני בפרסור מותאם של הדאטה, ולאחר מכן בזריקת פיצ'רים שנגנזו בתהליך הרגולריזציה של Lasso, ולבסוף ע"י מדידת הקורלציה בין פיצ'רים כדי לקבוע אם הם מתנהגים באופן קו-לינארי.
שיטות בהן השתמשנו:
לצורך model-selection ו-parameter-selection, השתמשנו בשיטת K-fold לcross-validation, עשינו זאת עבור כל מודל, ולכל מודל עשינו זאת עבור כל פרמטר בטווח הגיוני. קיבלנו score, ולקחנו את חמשת המודלים עם הפרמטרים שקיבלו את הscore הטוב ביותר (score ממוצע בשיטת k-fold). לאחר מכן, שלחנו את חמשת המודלים הטובים ביותר לבדיקה על הvalidation-set, ממנו קיבלנו את גרף הmse כתלות בגודל הדאטה שאימנו באמצעות המודלים הללו. מתוך חמשת הגרפים, בחרנו את המודל

שהגרף שלו הראה את הירידה ה"קלאסית" ביותר של השגיאה לאורך זמן, וזהו המודל אותו הפונקציה החזירה כמודל הנבחר.
התוצאות לאחר ביצוע תהליך זה באופן סופי הראו שהמודל הטוב ביותר עבור חיזוי הrevenue הינו מודל מסוג RandomForestRegressor עם 51 עצים, עומק מירבי 8, ומינימום samples לsplit של 6. כמו כן, המודל הטוב ביותר עבור חיזוי הvote-average הינו מודל מסוג GradientBoostingRegressor עם הפרמטרים הדיפולטיביים של פייתון עבור מודל זה.
להלן גרף לדוגמא של MSE vs. train_set_size, של המודל הטוב ביותר לחיזוי vote_average:



★ **Provide a prediction (and explanation) of the generalization error you expect your system to have?**

המערכת הלומדת שלנו כאשר נבדקה על test-set בגודל 1415 דגימות, נתנה שגיאה כוללת של 5e15 עבור חיזוי רווח ו-1.5 עבור חיזוי ממוצע ההצבעות. סה"כ, משמעות הדבר היא שגיאת הכללה ממוצעת של עד פי 10 טעות עבור חיזוי רווח, כלומר סדר הגודל הוא זהה עבור חיזוי רווח. ועבור חיזוי ממוצע הצבעות, הטעות קטנה עד מאוד.