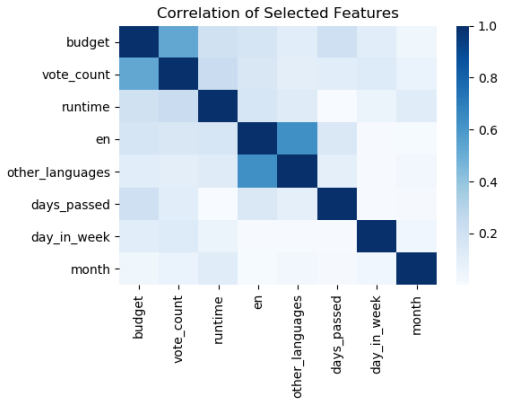
תהליך העבודה

- חילקנו את הדאטה ל-train(60%), validation(20%), test(20%)
- למדנו את ה-data לפי ה-train:
  - בדקנו עבור הפיצ'רים השונים מהם הערכים שנצפה לקבל. בכל עמודה קבענו את הפורמט, ואת הערכים שהגדרנו כערכים חוקיים.
  - הגדרנו את הטיפול במקרים של ערכים לא חוקיים, כדוגמת תאריך בפורמט לא תקין, או תאריך לא חוקי. בדקנו עבור כל עמודה שכיחות של הופעת ערכים שאינם חוקיים, וטיפלנו בהתאם - למשל קביעת ערכים סביב החציון.
- שלב הpreprocessing:

בשלב זה הפכנו את העמודות השונות לערכים עליהם אלגוריתמי הלמידה האופציונליים יוכלו לרוץ. בתוך כך:

- קבענו מספר משתנים קטגוריים. על מנת לא ליצור משתנים קטגוריים רבים מידי, בחלק מהמקרים איחדנו קטגוריות לפי התנהגות דומה, קורלציה גבוהה ביניהם, ולפי שכיחויות.
- במקרים מסוימים פיצלנו עמודה למספר עמודות שונות, חלקן רציפות וחלקן קטגוריאליות – לדוגמה, בעמודת הזמן מאז פרסום הסרט, הוספנו 3 עמודות שונות – יום בשבוע, חודש בשנה, ומספר הימים שעברו מאז פרסום הסרט.
- עבור ערכים חסרים ביצענו את הטיפול שהגדרנו לכל פיצ'ר בנפרד – החלפה לחציון, יצירת קטגוריה מתאימה לערך ריק וכו'
- בחירת הפיצ'רים:
  - במהלך שלב הpreprocess התקבלו פיצ'רים רבים. הדבר יכול להוביל לoverfitting ולבעיות נומריות, ולכן עבדנו על צמצום הפיצ'רים. העקרונות המנחים היה בחירת מספר פיצ'רים יחסית קטן עם קורלציה נמוכה ביניהם, ועם קורלציה גבוהה למול וקטור התוצאה, ושעבורם נקבל הטיה נמוכה.
  - כדי לבחור את תת הקבוצות של הפיצ'רים, השתמשנו בארבע שיטות:
    - בדיקת קורלציה באמצעות מתאם פירסון בין הפיצ'רים לבין עצמם כדי למצוא תלות

      בתרשים זה ניתן לראות למשל שעבור הvote_count והbudget, ישנה קורלציה יחסית גבוהה, אולם בחנו את ההטיה עבור מודלים שניסינו במקרה שבו מורידים את אחד מפיצ'רים אלו, ומאחר שההטיה גדלה באופן משמעותי, החלטנו להשאיר את שני הפיצ'רים.

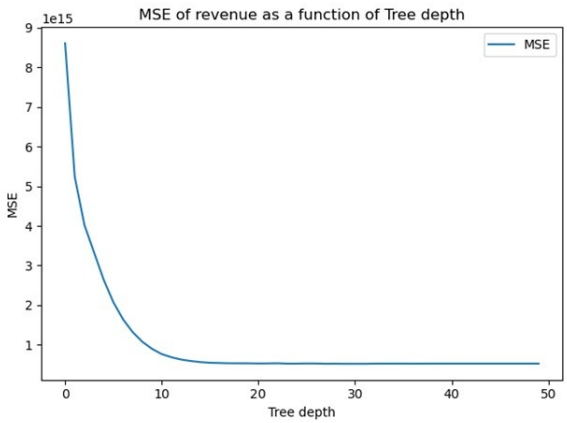


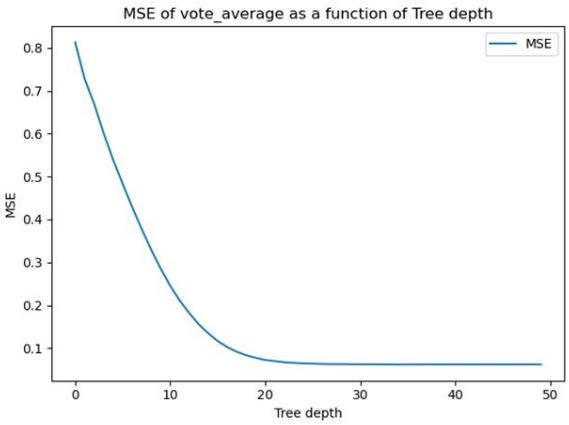
    - בדיקת הקורלציה באמצעות מתאם פירסון של הפיצ'רים מול ווקטור התוצאות.
    - בדיקת הקשר בין פיצ'רים לבין ווקטור התוצאות באמצעות המודל:

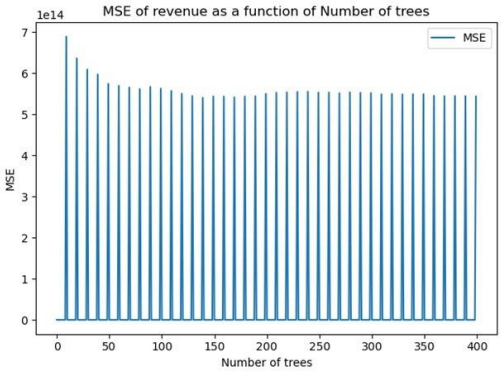
      sklearn.feature_selection.mutual_info_regression שמבוסס על k שכנים קרובים

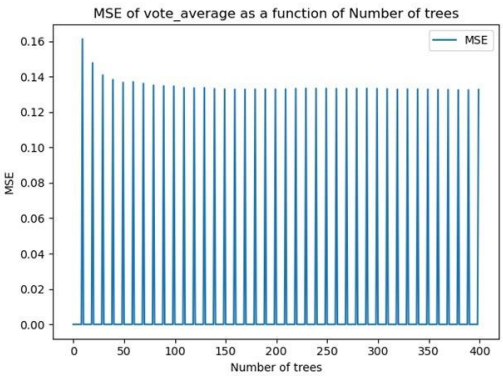
- השווינו את התוצאות שהתקבלו משלוש השיטות, ויצרנו תת קבוצות פיצ׳רים חשודים. לכל קבוצת פיצ׳רים השתמשנו ביער מקרי כדי להעריך את הביצועים.
  וכך בסופו של דבר בחרנו קבוצות עם מספר פיצ׳רים נמוך יותר כדי להוריד את השונות, ובו זמנית עם הטייה נמוכה.

- בחירת המודלים :
  - בדקנו מודלים שונים הdata שהתקבל משלב הpreprocess :
    - Adaboost שמתבסס על עצי רגרסיה
    - Lasso
    - Ridge
    - Random Forest
    - Linear Regression
  - את הרגרסיה הלינארית הגדרנו כ-baseline ועבור כל שאר המודלים, להם פרמטר רגולריזציה, הרצנו על ה-validation set והשוונו עבור כל מודל בין טווח פרמטרים אפשריים. להלן הבדיקות עבור Lasso, Ridg, Random Forest :









- לבסוף בחנו את הביצועים של כל המודלים, בהתאם לכל וקטור תוצאה שונה ובחרנו את המודלים עם הביצועים הטובים ביותר