המשימה בה בחרנו היא המשימה הראשונה - משימת הסרטים.

תיאור המידע:

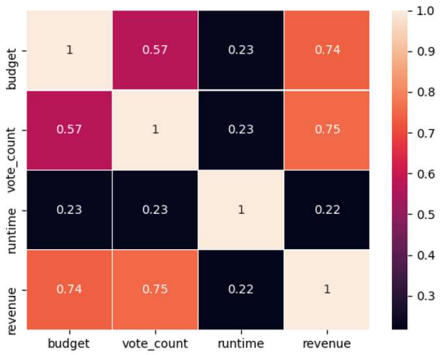
קיבלנו מידע עם עמודות רבות, אך עם מעט דגימות באופן יחסי. כמו כן, רוב מוחלט מהפיצ'רים היו טקסטואלים, רובם בפורמט JSON. בנוסף, ללא מעט מן הדגימות היה חסר מידע במאפיינים מסוימים. נציין שללא מעט דגימות היה חסר ה revenue עצמו מה שהופך את הדגימה לחסרת משמעות.

אתגרים באפיון המידע:

1. הבנה כיצד להמיר JSON לתצורה נוחה לעבודה כ-DATA-FRAME. על אחת כמה וכמה כיצד לעבוד עם JSON מקונן. חלק גדול מהזמן היה מושקע בפרסור הטקסט.
2. ניסיון להבין את תורת השימוש ב-free text לטובת בעיות למידה.
3. מילוי חוסרים במידעים של דגימות מסוימות (האם בכלל למלא את שורות אלו, אם כן, באיזה ערכים (ממוצע / מינימלי / חציון וכו')).
4. מידע קטגוריאלי רב ברירתי (לדוגמה - מספר רב של שחקנים) צריך להחליט אילו שחקנים הם המשמעותיים ביותר וכן לשאר הפיצ'רים בסגנון זה).
5. חלק מהפיצ'רים היו תלוים זה בזה, והיה צורך להבין איך והאם הם מיתרים אחד את השני (לדוגמא, מיקום של חברת הפקה והמקום בו צולם הסרט).

אופן ניקוי המידע (preprocessing):

1. הפיכת התאריך למשתנים אינדיקטיביים יותר: יום בשבוע, חודש ושנה.
2. ניקוי שורות שהסתמנו כרעש.
3. הפיכת משתנים רבים לקטגוריאלים, לדוגמה: companies, genres
4. בנוסף, את חלק מהמשתנים הקטגוריאלים איחדנו לאור תרומתם המעטה למודל. לדוגמה: רק חברות הפקה שמספר הסרטים שלהן בדאטה היה משמעותי, קיבלו קטגוריה משלהם (והקטנות אוחדו לקטגוריה אחרת) וכן ריכוז כל החודשים לפיצ'ר של יחיד של חודשים משמעותיים (חגים לדוגמה).



שיקולים שהנחו אותנו בעיצוב מודלי הלמידה:

1. מספר ניסויים ראשוניים בניסיון למצוא קורלציות (ללא מידע מקדים) ניתן לראות גרף מצורף.
2. בחינת מספר מודלים ליניאריים.

השיטות שניסינו והתוצאות שהן הניבו:

1. ניסינו להשתמש ב-bag of words גם לפיצ'ר overview וגם לפיצ'ר title. ניסויים אלה העלו חרס.
2. שימוש ב R.
3. כמו כן, לאור כמות המידע המועט, ניסינו להשתמש ב-ensemble. בסוף בחרנו להשתמש ב-random forest.
4. לאור כמות הפיצ'רים הגדולה שנוצרה (לאור הפיכת משתנים רבים ל-dummies), ניסינו להשתמש ב-feature selection, כדי להוריד את כמות הפיצ'רים.
   בשני הניסיונות, lasso ו-selectKBest לא ראינו תוצאות חד משמעיות.
5. על מנת למצוא את המודל הטוב ביותר, ניסינו להשתמש במודלים הבאים: Random forest, Regression, Gradient boosting. לבסוף, החלטנו לדבוק Random forest מאחר שהוא נתן לנו את התוצאות הטובות ביותר.

שגיאת ההכללה של החיזוי של המודל שלנו:

במהלך האימון הגענו ל-R^2 של כ- 0.8. ול RMSE של 62782878.193. (על הטסט שלא נגענו בו במהלך כל האיימון כלל כמובן.) אנחנו חושדים שהיה לנו overfit על הtrain-set והיינו רוצים לשנות את הפרמטרים של ה random-forest כדי לטפל בכך,אולם לא הספקנו להגיע לזה.

לסיכום, רוב ההתעסקות שלנו היתה עם הדאטה, וחילוץ המידע ממנו היה מאוד מאתגר, עקב כך לדעתנו לא הספקנו להתעסק מספיק עם המודל עצמו ולשפר אותו.