IML HACKATHON

אהוד דהן, נתן גולדשטיין, תמר עשהאל, נריה כהן

IML_HACKATHON_2021

Aim We strive to create a model that, given data about movies, will successfully predict the revenue made in the box office and the average viewer ranking of those movies before they are officially released. The revenue is represented as an int, and the ranking is a float between 0 and 10, with one digit after the decimal point.

DATASET DESCRIPTION AND CHALLENGING CHARACTERISTICS:

Dataset description: The provided dataset contained around 5,000 movies, each with 22 features (ID, collection to which the movie belongs (if indeed), budget, genre, link to homepage, original language, original title, overview, number of viewers who ranked the movie, production companies, production countries, release date, runtime, spoken language, the stage of production, tagline, title, keywords, cast, crew.

Challenging Characteristics of the Dataset:

1. The multitude of formats that the different characteristics of the movies had. Some of the data (like the budget) was numeric, some was a long string (like the overview). Moreover, some columns included lists inside of them: the genres, crew, production company, keywords, cast and spoken languages, were all lists.
2. Nonnumeric data had to be processed before it could be used.
3. Missing data. Many values simply did not exist in the provided data. For example, many revenue values were missing.
4. Non-useful variables. Some of the features could be directly inferred from other features and were therefor pretty much useless in and of themselves.

DATA CLEANING AND PRE-PROCESSING

First,

we had to figure out a way to divide the provided data into a training set and a set for testing. We decided to do this by first sorting the all the data according to the revenue, and then picking every forth feature to be in the test set, the rest of the data went to the training set. We decided to fill missing fields using the average of the data set, thus we would have a better result than if we were simply to fill the missing space with a zero, and we do not change the balance of that feature too much. We dealt with nonnumeric data by using dummy variables. Now, we feel that since much of the course is in English, it is appropriate to write at least some of the document in Hebrew. So -

CONSIDERATIONS THAT GUIDED OUR DESIGN OF THE LEARNING SYSTEMS and DESCRIPTION OF THE CHOSEN ALGORITHM

חשבנו בהתחלה להשתמש במודל רנדום פורסט אבל החלטנו לרדת מזה עקב כך שהיה נראה שההתפלגות של הדאטא היא יחסית ליניארית ובעיקר אלכסונית ואילו מודל רנדום פורסט צריך עומק גדול מאוד בשביל לייצר הפרדה אלכסונית, וחששנו שזה יוביל לאובר פיטינג. אפשרות נוספת היתה להשתמש בשכנים קרובים, זאת היתה יכולה להיות החלטה נבונה, אבל חשבנו שנשתמש במתודה זו בעיקר בשביל להשלים ערכים חסרים בדאטא. ואם היינו משתמשים בה גם בשביל ערכים חסרים וגם בשביל לנבא את הדאטא הכללי היינו מקבלים אובר פיטינג. בנוסף לכל אלה, ממטריצות הקורלציה היה נראה שיש משתנים שמסבירים את הדאטא באופן ליניארי די מובהק)כגון התקציב, כמות ההצבעות, וכו(ולכן חשבנו שהדאטא בסהכ מתנהג באופן ליניארי ברובו ומכאן נבעה החלטתנו להשתמש במודל רגרסיה לינארית.

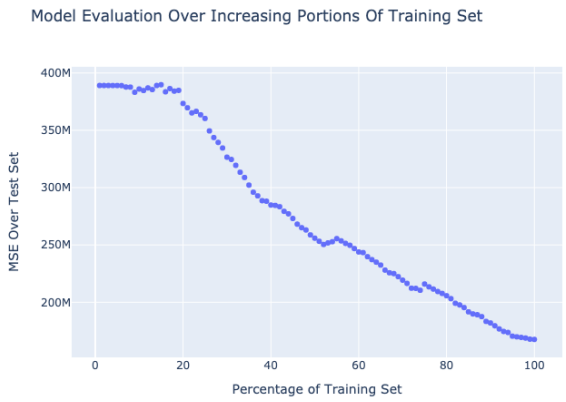
VARIOUS METHODS WE TRIED AND THEIR RESULTS

על מנת לדלל את כמות הפיצ׳רים נקטנו במספר שיטות. ראשית עשינו plot scree של הערכים העצמיים כפי שלמדנו בתרגיל 2. לאחר מכן הפעלנו מתודת לאסו על מנת למצוא פיצ׳רים שמתאפסים. בנוסף יצרנו מטריצת קורלציות על מנת למצוא משתנים תלויים, משתנים חשובים לניבוי, משתנים חסרי משמשעות לחלוטין וכו.)מצורפים תרשימים רלוונטים(בסופו של דבר עשינו בחירה ידנית של משתנים שחיבור של המתודות אלו יחד הsense common שלנו חשבנו שהם יהיו המשמעותיים ביותר לניבוי.

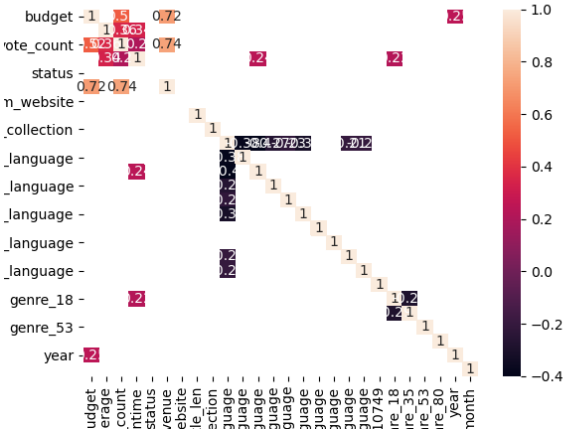
PREDICTION (AND EXPLANATION) OF THE GENERAL MODEL ERROR WE EXPECT OUR SYSTEM TO HAVE:

אנחנו מצפים שככל שנקבל דאטא יותר מורעש/חסר נתונים יהיה עלינו קשה יותר לנבא נכון והטעות שלנו תגדל.

הפלוטים בעזרתם קיבלנו החלטות:



מטריצת קורלציות


budget
ote_count
status
n_website
_collection
_language
_language
_language
_language
_language
genre_18
genre_53
year
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4