**Box Office Revenue Prediction**

מגישים:אריאל ברגר ויונתן קודלש

ת"ז: 315472084, 203981386

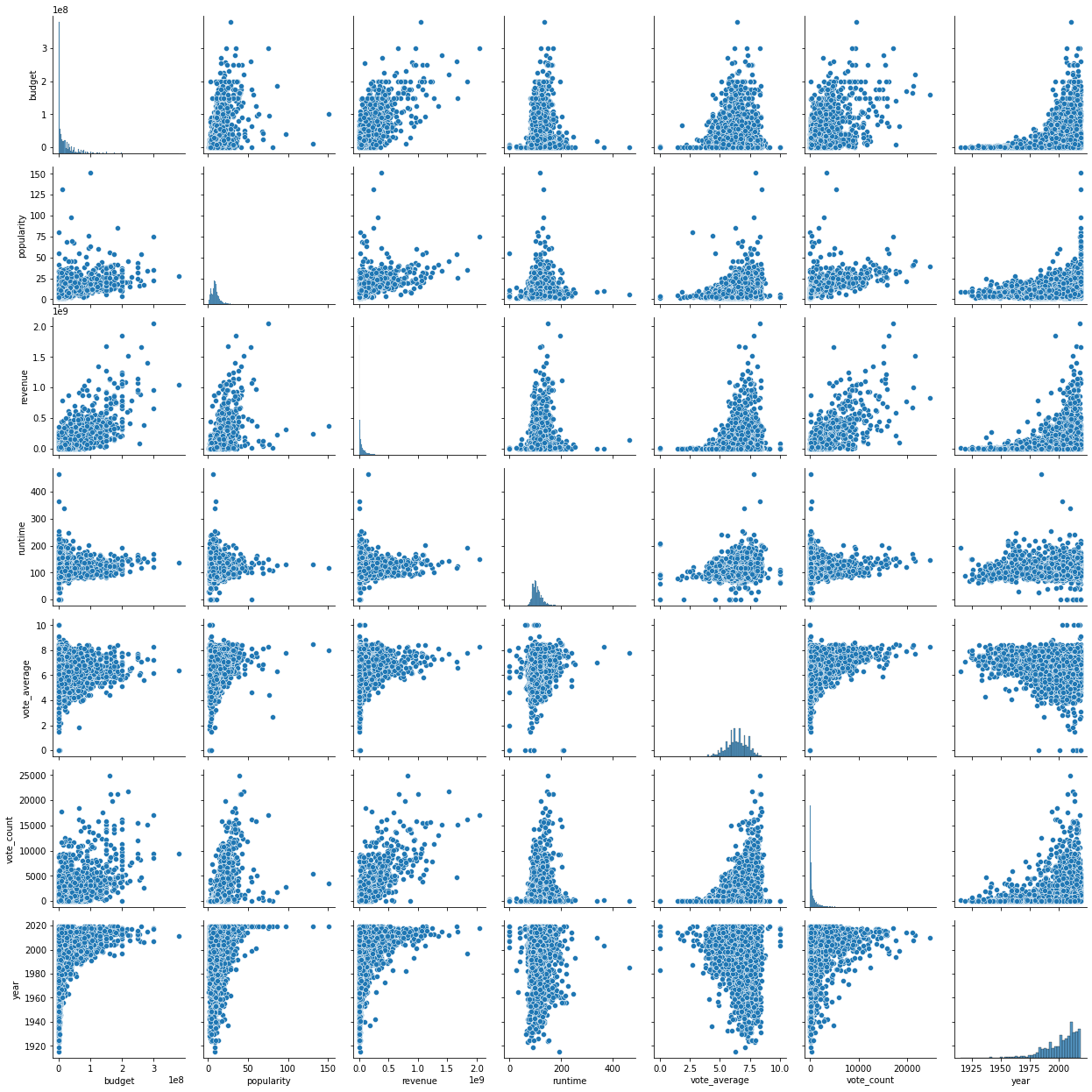
**Exploratory Data Analysis**

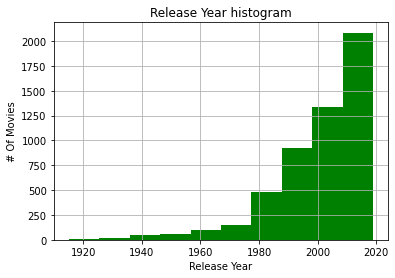
1. הפיצ'רים  
   בדאטה סט ישנם סך הכל 26 פיצ'רים המתארים מאפיינים של סרטים, כאשר כל רשומה מתארת סרט. משתנה המטרה הינו revenue - סך המכירות. באופן כללי ניתן לחלק את הפיצ'רים ל4 קבוצות עיקריות:
   1. פיצ'רים נומינלים (budget, vote count, popularity..)
   2. פיצ'רים קטגוריאלים (cast, crew,belongs to collection,...), בהם כל רשומה מכילה json עם מספר פרטים הרלוונטים לפיצ'ר - למשל העמודה cast מכילה json עם כל שמות השחקנים ושמות הדמויות.
   3. פיצ'רים נוספרים המתארים את הסרט אך אינם רלוונטים למשימה - (backdrop path, imdb id, poster path,...).
   4. פיצ'רים שבהם מופיע מידע בצורת טקסט (overview, tagline) .

בנוסף ישנו פיצ'ר אורדינלי המציין את תאריך היציאה של הסרט.

1. השוואה בין פיצ'רים

* ראשית, אנו מציגים בגרף 1 מתחת, את הקשר בין כל זוג משתנים נומינלים באמצעות scatter plot. בנוסף, ישנה היסטוגרמה של כל אחד מהפיצ'רים בפני עצמו. ניתן להבחין במספר דברים מעניינים באמצעות הגרף:
* כאשר מתמקדים בשורה הרלוונטית לrevenue רואים כי לכל המשתנים הנומינלים נראה שישנו קשר עם מתאם כללי חיובי בינם לבין ההכנסה (משתנה המטרה).
* בולטים במיוחד בהקשר זה הם המשתנים budget ושנת היציאה של הסרט. ניתן לראות כמעט יחס ישיר בין עלייה בתקציב לעליה בהכנסות \ שנת יציאה.
* רוב המשתנים מתפלגים נורמלית על הruntime.
* נראה כי משתנה המטרה revenue מתפלג פרטו עם זנב ארוך מאוד.



גרף 1 - קשר בין משתנים נומינליים

* המשתנה video (בוליאני) מכיל אך ורק 5 סרטים בעלי הערך True בסט האימון
* 19% מהסרטים מכילים הם חלק מסדרות סרטים.
* רובם המוחלט של הסרטים בסט האימון יצאו בשנים האחרונות (מוצג בגרף 2)

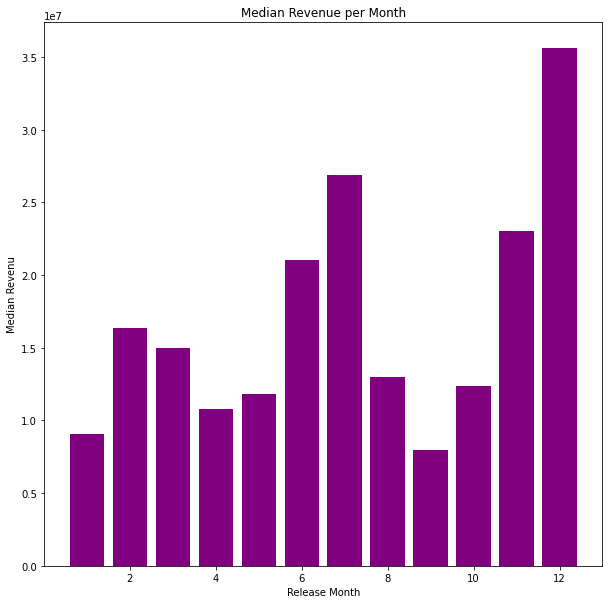
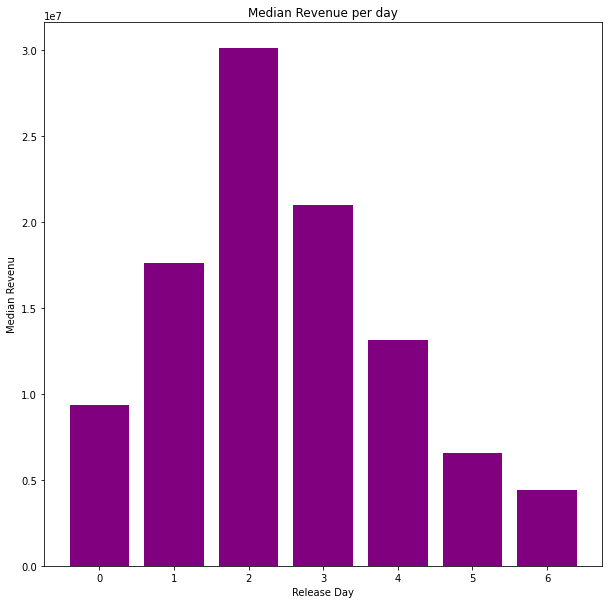
גרף 2 -היסטוגרמת שנת יציאת הסרטים

1. ערכים חסרים:

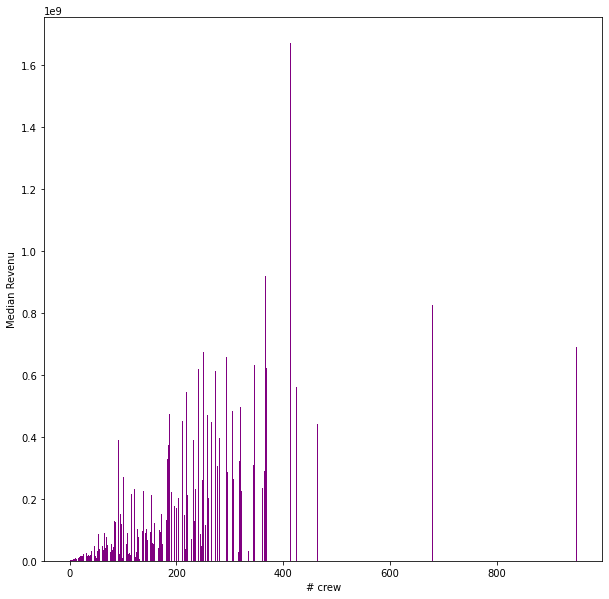
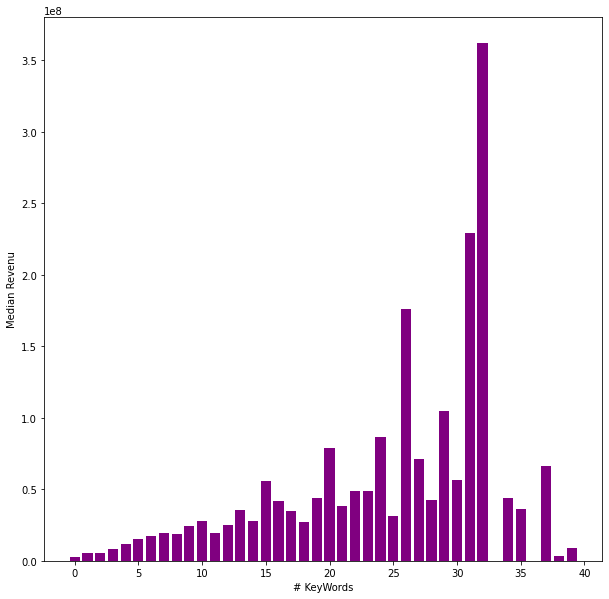
* בעמודות backdrop path, homepage, runtime, belongs to collection, poster path, imdb id & tagline ישנם ערכים חסרים. מתוך עמודות אלו השתמשנו אך ורק ב belongs to collection (שבעבורה ניגע בהמשך איך ביצענו את הנדסת הפיצ'ר).
* המשתנה budget אמנם אינו בעל ערכים חסרים, אך נראה שחלה טעות בעת בניית הדאטה סט. בקובץ האימון, 29% מהרשומות מכילות תקציב 0. זוהי כנראה טעות. בחלק הבא ניגע ונסביר איך ניסינו לטפל בדאטה הפגום.
* למשתנה המטרה revenue (סך ההכנסות) ישנם מעט ערכים נמוכים מאוד (3% מהרשומות מכילים הכנסות של מתחת ל10K$). ככל הנראה התרחשה כאן טעות (חיפוש קצר בגוגל מאשש את הטענה הזאת).

**Feature Engineering**

1. **הפיצ'רים בהם אנו משתמשים:**

* פיצ'רים הקשורים לתאריך שחרור הסרט - שנת היציאה, החודש והיום בשבוע שבו הסרט. בחלק הראשון, ניתן לראות את המתאים החזק בין שנת היציאה של הסרט לבין סך ההכנסות (revenue). בנוסף, בדקנו ומצאנו כי סך ההכנסות משתנה אף בתלות בחודש שבו הסרט יצאה וביום בשבוע כפי שניתן לראות בגרפים 3 ו4.

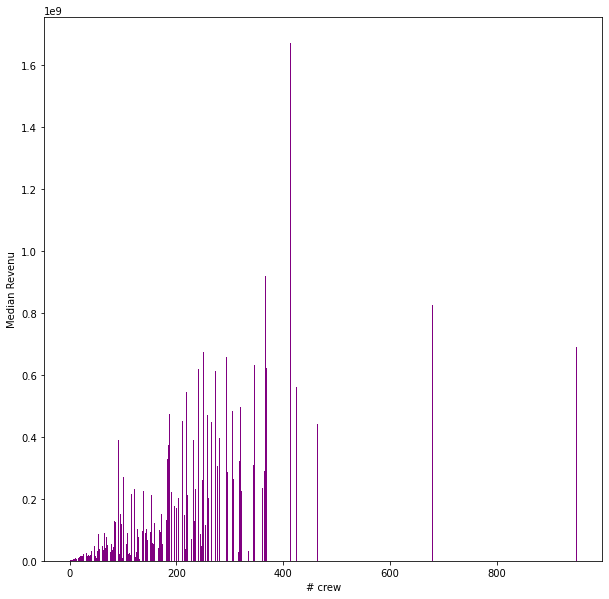
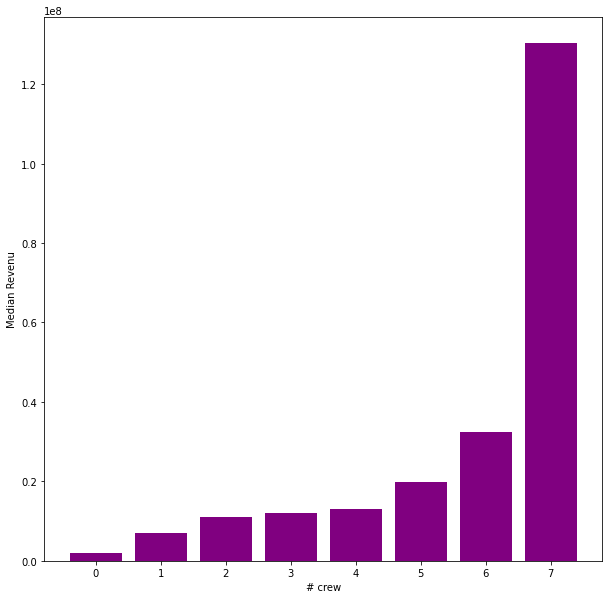
גרפים 3 ו-4 - הכנסה חציונית כתלות ביום בשבוע בו יצא הסרט (ימין) \ בחודש שבו יצא הסרט

* פיצ'רים נומינליים - כפי שראינו בחלק הקודם בגרף 1, לפיצ'רים popularity, vote count, vote average, budget ישנה קורלציה ישירה לסך ההכנסות והם אכן עזרו לנו לחזות ובחרנו להשתמש בהם. לגבי הפיצ'ר runtime, ראינו אמפירית כי ללא פיצ'ר זה התוצאות של המודל שלנו טובות יותר ולכן החלטנו להוריד אותו. עובדה זו מתחברת לכך שראינו כי הruntime מתפלג נורמלית בשאר הפיצ'רים הנומינליים, ייתכן שישנה פה אי תלות.
* הפיצ'רים הבאים ראינו כי משפרים את תוצאות המודלים בצורה אמפירית:
  + בפיצ'ר genre
  + פיצ'ר המציין האם ארץ המוצא של אחת מחברות ההפקה של הסרט נמצאת ב10 הארצות עם המספר הרב ביותר של חברות סרטים (10 נבחר על ידי קרוס וואלידציה)
  + פיצ'ר בוליאני is top prod המציין האם הסרט מיוצר על ידי אחת מ10 חברות ההפקה הגדולות ביותר המופיעות בסט האימון. (10 נבחר על ידי קרוס וואלידציה)
  + פיצ'ר has collections המציין האם הסרט הוא חלק מסדרה.
  + פיצ'ר is en - המציין האם שפת המקור של הסרט היא אנגלית.
  + הפיצ'רים spoken languages, cast, crew, Keywords - פיצ'רים נומינלים הסופרים את כמות השפות \ אנשי צוות \ שחקנים \ מילות מפתח הקשורות לסרט. בגרפים 5,6 ניתן לראות את החשיבות של הפיצ'רים Keywords ו - crew על בסיס ההכנסה החציוני ביחס כל אחד מהשדות.

**גרפים 5 ו-6 - הכנסה חציונית כתלות מספר מילות המפתח בסרט(ימין) \ במספר אנשי הצוות (שמאל)**

* פיצ'רים מעניינים נוספים שניסינו להכניס למודל אך לא עזרו לשפר את הביצועים ולכן הורדנו אותם - יחס בין התקציב לשנה \ פופולריות לשנה \ תקציב למספר הצבעות \ פופולריות למספר הצבעות \ האם מופיע שחקן מהשחקנים הפופולרים ביותר (גם בינארי וגם one hot encoded לטובים ביותר) \ יחס בין התקציב למספר השחקנים או הצוות ועוד.

1. **Feature transformation**

* את הפיצ'רים belongs to collections / original langauge העברנו לאינדיקטור בוליאני הבודק אם קיימת רשומה \ האם השפה היא אנגלית.
* לשם הייצוג החדש של התאריך, ביצענו one hot encoding לחודש היציאה והיום בשבוע בו הסרט יצא, בנוסף לפיצ'ר המכיל את שנת יציאת הסרט.
* את הפיצ'ר is top prod יצרנו באמצעות השדה production companies. בדקנו מי הם 10 החברות הגדולות ביותר, ועברנו על כל רשומה ובדקנו האם אחת החברות שיצרה את הסרט היא מחברות ההפקה הגדולות.
* באופן דומה יצרנו פיצ'ר קטגוריאלי בעבור 10 מדינות המוצא של הפקות הסרטים הנפוצות ביותר. לפיצ'ר זה ביצענו בנוסף one hot encoding.
* לפיצ'רים spoken languages, cast, crew, Keyword - כאמור חילצנו מהם את כמות השפות \ אנשי צוות \ שחקנים \ מילות מפתח הקשורות לסרט. כדי לשפר את ההכללה של המודל, ביצענו ניסויים בהם המרנו את הערכים הכמותיים לפיצ'רים אלו ל8 binns בגודל שווה. הבינניזציה עזרה לשפר את תוצאות המודלים (המספר 8 גם כן מבחר אמפירית). בגרפים 7,8, המציגים את ההכנסה החציונית בתלות בפיצ'ר crew, לפני ואחרי פעולת הbin ניתן לראות כי לאחר הטרנספורמציה הפיצ'ר פחות רגיש לרעש ואינדיקטיבי יותר:

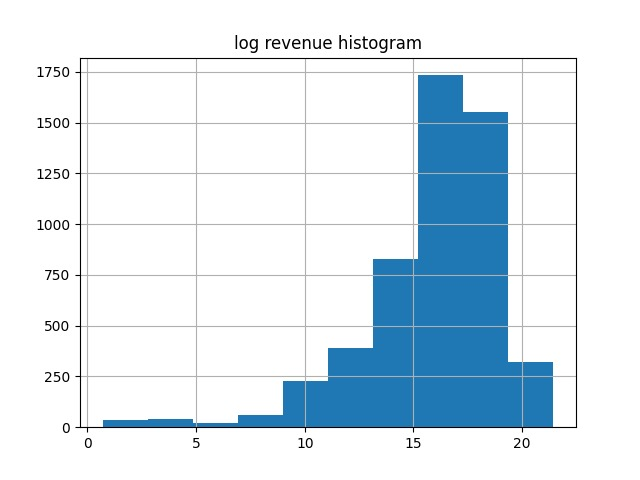
גרפים 7 ו-8 - הכנסה חציונית כתלות במספר אנשי הצוות לפני (שמאל) ואחרי פעולת הbin (ימין)

1. **Handling missing data**

* כאמור בחלק הEDA, למשתנה התקציב (budget) ישנם ערכים שגויים בסט האימון (סרטים עם תקציב 0). בדקנו וראינו כי מרבית הרשומות הללו, בעלי תקציב 0, ישנם גם סך הכנסות נמוך. מכיוון שיש קשר חזק בין התקציב לבין ההכנסות, לקחנו זאת בחשבון כאשר ניסינו מספר שיטות imputation שונות על מנת לטפל ברשומות עם הערך 0.

הטכניקות השונות בהן השתמשנו - השלמה לחציון או אחוזונים שונים של התקציב \ רגרסיה באמצעות שאר הפיצ'רים \ חציון לפי העשור שבו הסרט יצא (בגרף 1 ראינו את הקשר המשמעותי בין תקציב לשנה שבה הסרט יצא).  
לבסוף, לאחר קרוס וואלידציה על מספר מודלים וביצועיהם עם ערכי הbudget החדשים, השיטה שהתגלתה אמפירית כטובה ביותר היא השלמה לאחוזון ה10 של התקציב בסט האימון. להפתעתנו, שיטה פשוטה זו התעלתה על השיטות המתוחכמות יותר בהן ניסינו לשערך את התקציב ברשומות אלו.

* בסעיף הEDA אף הזכרנו כי למשתנה המטרה ישנם מספרים ערכים שנראה חריג ולא הגיוני (הכנסה מאוד נמוכה). בגרף 9, המציג היסטוגרמה של לוג ההכנסות של סט האימון, ניתן לשים לב לערכים אלו בחלק השמאלי של הגרף. כדי לנסות ולתקל את הבעיה הנ"ל, ניסינו להוריד את 2.5% הסרטים עם ההכנסה הנמוכה ביותר בזמן האימון. לצערנו, באמצעות הקרוס וואלידציה ראינו כי ביצועי המודל פוחתים בעקבות שינוי זה ולכן השארנו את הערכים כפי שהם במקור.



גרף 9 - היסטוגרמת לוג ההכנסות

**Prediction**

בשלב הראשוני, השתמשנו בחבילת [pycaret](https://pycaret.org/), המכילה מספר רב של מודלי ML. באמצעותה בדקנו מי הם המודלים המבטיחים ביותר מבין כל המודלים, על מנת שנוכל להמשיך להעמיק ולמצוא בעבורם את ההיפרמטרים האידיאליים. כמצופה, מודלים המבוססים על Gradient Boosting קיבלו את התוצאות הטובות ביותר. מבין מודלי הML, התמקדנו ב2 - XGBoost ו lightGBM.

בנוסף למודלים אלו, לשם השוואה ועניין, חקרנו וניסינו לחזות את ההכנסות באמצעות למידה עמוקה ורשת נוירונים.

כעת, נעבור על המודלים השונים, כאשר את התוצאות נציג ונשווה בטבלה 4 למטה:

1. **מודל XGBoost**

זהו מודל גרדיאנט בוסטינג קלאסי, המשתמש בשיטות בוסטינג ומכיל מספר רב של של weak learners ומשתמש בשיטות גרדיאנט על מנת למצוא את העצים האידיאלים.

אימנו את המודל באמצעות חבילת [optuna](https://optuna.org/), המוצאת את ההיפר פרמטרים האידיאלים לכל מודל באמצעות שילוב של אלגוריתמי grid search, random, bayesian, & evolutionary algorithms למציאת היפר פרמטרים אידיאלים. על מנת לקבל הערכה טובה של ביצועי המודל עם הפרמטרים השונים השתמשנו בחלוקה של הדאטה 10 fold והערכה של כל סט היפר פרמטרים לפי ממוצע הRMSLE.

ההיפר-פרמטרים שנבחרו:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| max depth | colsample bytree | subsample | alpha | lambda | booster |
| No max | 0.4771 | 0.7520 | 6.4953 | 0.2719 | gblinear |

טבלה 1 - היפר פרמטרים נבחרים של מודל ה**xgboost**

1. **מודל LightGBM**

זהו מודל גרדיאנט בוסטינג, המשתמש בשיטות בוסטינג ומכיל מספר רב של של weak learners ומשתמש בשיטות גרדיאנט על מנת למצוא את העצים האידיאלים. להבדיל מXGBoost, מודל זה משתמש בשיטות מתוחכמות יותר על מנת לבצע את שיטות הsampling. שיטות אלו הופכים אותו ליעיל הרבה יותר מבחינת זמן ריצה.

במודל זה גם כן השתמשנו בoptuna וב10 fold validation על מנת למצוא את ההיפר-פרמטרים האידיאלים.

ההיפר-פרמטרים שנבחרו:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| max depth | feature fraction | bagging fraction | min data in leaf | num leaves | learning rate | boosting type | |
| No max | 0.4 | 0.96 | 25 | 8 | 0.1 | gbdt |

טבלה 2 - היפר פרמטרים נבחרים של מודלה**LightGBM**

1. **מודל Neural Network**

על אף שהשימוש המרכזי ב deep learning אינו מיועד בדרך כלל למידע טבלאי, ניסינו לאמן רשת FC פשוטה עם מספר שכבות חבויות עם אקטיבציה בינהן על מנת להשוות לאלגוריתמי הגרדיאנט בוסטינג.

גם במקרה זה השתמשנו בחבילת optuna על מנת למצוא את ההיפרפרמטרים האידיאלים של הרשת. היפר הפרמטרים הנבחרים הם:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Dropout 2 | Layer 2 hidden dim | Dropout 1 | Layer 1 hidden dim | epoch | Optimizer | lr | # hidden layers |
| 0.13 | 34 | 0.26 | 861 | 10 | SGD | 0.034 | 2 |

טבלה 3 - היפר פרמטרים נבחרים של מודלה**NN**

תוצאות

לאחר בחירת היפר הפרמטרים אימנו את המודל על מלוא הדאטה. בטבלה 4 לפניכם אנו מציגים את התוצאות. נבחין במספר נקודות:

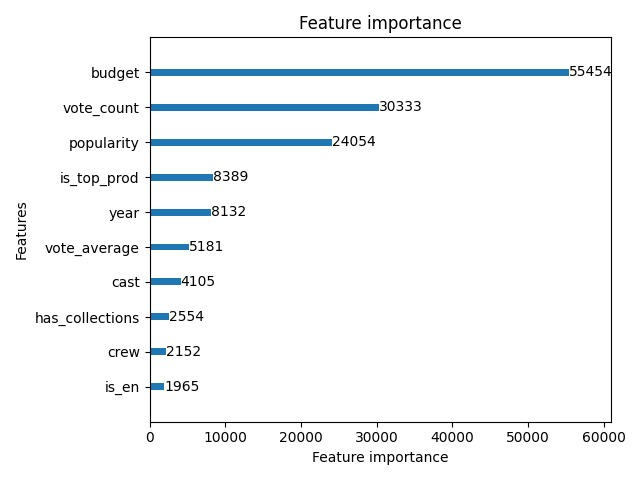
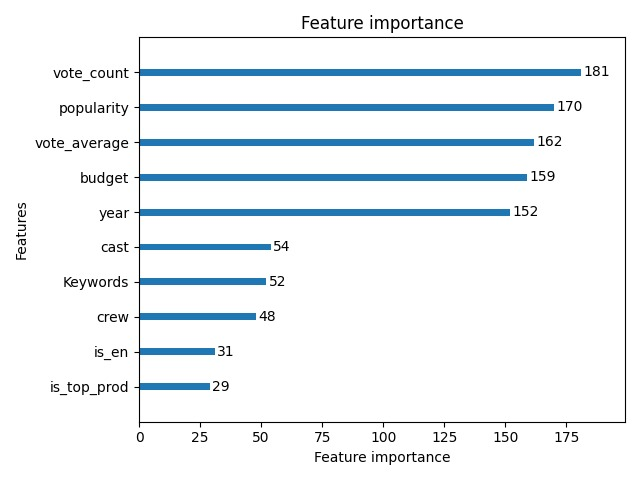
* ביצועי מודלי הגרדיאנט בוסטינג התעלו בהרבה בהשוואה למודל הרשת העמוקה.
* ביצועי כל המודלים על הטסט היו טובים יותר בהשוואה ל validation. אנו מייחסים שיפור זה לכך שאת המודלים אימנו מחדש על הדאטה המלא ולא רק על 90% מסט האימון, עם היפר הפרמטרים האידיאלים שמצאנו.
* מודל הXGB היה המוצלח ביותר מבחינת RMSLE, אך מודל הLightGBM הגיע לביצועים דומים מאוד, על אף זמן אימון קצר בהרבה (סדר גודל של פי 4 יותר קצר).
* בסוף תהליך האימון, הרכבנו מודל אנסמבל שלוקח את ממוצע התחזית של מודלי הXGB ומודל הLightGBM. אנסמבל זה הצליח לשפר את התוצאות ולהגיע ל1.81 RMSLE על סט המבחן.

|  |  |  |
| --- | --- | --- |
| Model | mean Validation RMSLE | Test RMSLE |
| XGB | 2.0666 | **1.8659** |
| LightGBM | 1.9598 | 1.8698 |
| Neural Network | 2.3717 | 2.3572 |

טבלה 4 - תוצאות כל המודלים

4. בונוס - חשיבות הפיצ'רים

באמצעות אלגוריתם הLightGBM ניתן להוסיף explainability למודל כדי להבין את חשיבות כל אחד מהפיצ'רים, הן מבחינת ה information gain והן מבחינת כמות השימושים בפיצ'ר.

בגרפים 10 ו11 אנו מציגים את חשיבות הפיצ'רים לפי 2 הקריטריונים. לפי שני הקריטריונים, הפיצ'רים החשובים ביותר הם vote count, budget & popularity.

גרפים 10 ו11 - חשיבות הפיצ'רים על פי כמות ה**information gain** (שמאל) \ מספר הפעמים שהפיצ'ר נבחר לפיצול (ימין)