ניתוח הפרויקט - Project Analyze

המידע שלפנינו מכיל רשימה של סרטי וידאו, כשכל אחד מהסרטים מכיל מספר מאפיינים המגדירים אותו, כמו צוות השחקנים (crew), כמות השקעה כספית (budget) וכמות הרווח והציון שקיבלו.

המידע עצמו כלל מספר אתגרים: ראשית, פירוק המידע הבסיסי היה מורכב כשלעצמו – המידע מורכב מקבצי json ומעבר לכך ממערכים של קבצי json וחלקם אף הובאו עם מפרידים שבלתי אפשרי להתעסק איתם. קבצים אלו אינם ניתנים לעבודה בסיסית ומחולקים למספר פרמטרים שרבים מהם חוזרים על עצמם (למשל, id ושם שחקן – משתנה כפול). בעקבות כך, נדרשנו לכתוב מספר פונקציות שיהפכו את המידע לקל יותר לעבודה – בחרנו במשתנה אחד בכל פעם (למשל name) והפכנו את כל אחד מהתאים ל-dataframe. דבר זה אפשר לנו לנתח את המידע בשלב מאוחר יודע – כך יכולנו למשל לקחת את רשימת השחקנים הטובים ביותר ועוד.

מעבר לכך, בתוכן המידע עצמו היו מספר אתגרים. כיצד נוכל להשוות בין הפרמטרים השונים? מה הדרך הקלה לעבוד עם מידע ככזה? מה קורה במקרה בו אחד הפרמטרים הינו ריק, כיצד נוכל להתמודד עם מידע חדש שהגיע לנו ואיננו נמצא במדגם המקורי (בtraining data)? את אתגרים אלו פתרנו כל מקרה לגופו – כלומר, ניתחנו את המידע בהתאם לנתונים האחרים.

בתהליך המידע בחרנו למעשה מספר אסטרטגיות משמעותיות, שניתן לחלקן למספר סוגים:

1. ניתחנו אילו חלקי מידע חסרי קורלציה או חסרי יכולת פרדיקציה. לדוגמא, לא ניתן לחזות על סמך שם הסרט, ובעקבות כך מחקנו פיצ'ר זה מהמידע שלפנינו.
2. בדקנו כיצד הפיצ'רים מתמודדים במקרה בו יש מידע חסר – למשל במקרה בו ה-budget הינו 0, נקבל כי המידע שלנו חסר והפרדיקציה על פיו איננה אמינה.
3. ניתחנו אילו מהפיצ'רים תלויים אחד בשני בצורה מובהקת – דבר זה מיותר ואף עלול לפגוע בדיוק המידע ולכן בחרנו למחוק אחד מהפיצ'רים במידה והם תלויים אחד בשני. למשל, מחקנו את הproduction_country כיוון שהוא כבר מתקשר לproduction_company וכו'.
4. ניתחנו אילו פיצ'רים נוכל להמיר לone hot encoding ובאיזה צורה. בנקודה זו עבדנו לפי שתי גישות מרכזיות:
   a. בחרנו מספר נקודות סף (thresholds), לפיהן חילקנו את המידע לקבוצות, בהן מתקיים שינוי משמעותי. למשל, בעקבות בדיקת קורלציה שמצאנו, גילינו כי ניתן לחלק את זמן הסרט (runtime) לסרטים קצרים (עד 60 דקות), סרטים ממוצעים (עד 160) וסרטים ארוכים (160 ומעלה).
   b. בחרנו מספר פרמטרים שהינם דומיננטיים יותר משאר הפרמטרים. כלומר, משקלנו את הפרמטרים בחלק מהפיצ'רים, שהינם משמעותיים יותר משאר הפרמטרים. למשל, בחרנו שחקנים שמופיעים במספר רב יותר של סרטים והוספנו אותם ל-data שלפנינו – כך נוכל ליצור הסבר טוב יותר לפרדיקציה בעתיד.

לאחר שביצענו את תהליך הכנת המידע (preprocessing), נדרשנו לבחור מערכת למידה מתאימה. האפשרויות שהיו לפנינו היו מודל רגרסיה ליניארית פשוט, ועץ רגרסיה (Regression Tree). מודל הרגרסיה היה פשוט מדי, בעל שונות (variance) והטייה (bias) גבוהים, ולכן בחרנו את עץ הרגרסיה. לאחר מכן, מכיוון

שלא היה לנו המון מידע, רצינו לבצע boost. שימוש בadaboost לא התקיים בצורה מוצלחת, ולכן החלטנו להשתמש בrandom forest שמזער את שגיאת ההכללה בצורה טובה.

לאחר מכן, על מנת לקבוע מהם הפרמטרים הטובים ביותר (למשל, כיצד לבחור את גובה העץ), הרצנו את המודל על מספר קלטים שונים ולבסוף בחרנו את הפרמטר האידיאלי שעבורו נקבל הטיה ושונות אידיאליים.

לאחר כל השלבים הללו, הפעלנו מודל שמשקלל את כלל המודלים ומתייחסים לפרמטרים ספציפיים, למשל עבור מקרה בו הסרט עדיין לא יצא.

שגיאת ההכללה שקיבלנו עבור הרווח (revenue) הינה בערך 65 מיליון ועבור הדירוג הממוצע 0.7 כפי שצפינו, כיוון שחלק מהמידע שבידינו חסר, וכי עלינו לקבל מידע נוסף או לערוך מחקרים מעמיקים לגבי ההשפעה של חלק מהפיצ׳רים על המידע.