**Readme- Google Play Store Apps**

בפרויקט זה ניסיתי לחזות את הדירוג הממוצע של אפליקציה מתוך דאטה שמכיל 32,000 אפליקציות מחנות האפליקציות של גוגל.

הפיצ'רים שהיו בדאטה הם:

* שם האפליקציה
* לאיזה קטגוריה האפליקציה משתייכת
* מה הדירוג הממוצע שמשתמשים דירגו את האפליקציה
* כמה פעמים האפליקציה דורגה
* כמה פעמים הורידו את האפליקציה
* מה גודל האפליקציה
* מה מחיר האפליקציה בדולרים
* מה גיל הקהל המיועד לאפליקציה
* התאריך בו התבצע העדכון האחרון של האפליקציה
* מה היא גרסת האנדרואיד המינימלית המאפשרת שימוש באפליקציה
* מה היא הגרסה הנוכחית של האפליקציה

הכנת הדאטה:

העבודה העיקרית על הדאטה הייתה התאמת כל עמודה לביצוע החיזוי.

התחלתי את הסידור בבדיקת כל הקטגוריות הקיימות, גיליתי כי קיימות 49 קטגוריות כאשר 16 מהן הן קטגוריות הקשורות למשחקים. איחדתי את כל קטגוריות המשחקים לקטגוריה אחת כדי לקבל דאטה מסודר יותר, ונתתי לכל קטגוריה מספר שייצג אותה. כך קיבלתי קטגוריות הממוספרות מ- 1 עד 33.

לאחר מכן הפכתי את עמודת הרייטינג לעמודה מסוג מספרי (במקום אובייקט), והשתמשתי בפקודה שמעגלת כל מספר, כדי שלאחר מכן בשלב החיזוי יהיה אפשר לקבל מספר קטן יותר של ערכים.

לאחר הסתכלות על עמודת המחיר ראיתי כי רוב האפליקציות (30,000 כמעט) הן בחינם ורק מספר קטן יחסית (קצת מעל 2000) עולות כסף. לכן בעמודה זאת שיניתי כל אפליקציה שלא עלתה כסף לספרה 0, ואת שאר האפליקציות לספרה 1.

שיניתי את עמודת ההורדות כך שהערכים בה יהיו מסוג מספרי, לכן מספר ההורדות בכל אפליקציה משקף את מספר ההורדות המינימלי.

לאחר מכן הוספתי עמודה חדשה המייצגת את אורך השם של כל אפליקציה. הנחתי כי אורך השם של אפליקציה מסוימת יכול להשפיע על הדירוג של המשתמש גם אם ההשפעה מתבצעת באופן לא מודע. זאת מכיוון שאין למשתמש יכולת להוריד ולדרג אפליקציה בלי לראות את השם שלה.

לפני ביצוע התחזיות הורדתי את העמודות- שם אפליקציה, גודל אפליקציה, גיל קהל היעד, עדכון אחרון, גרסה מינימלית וגרסה אחרונה. זאת מכיוון שאני מאמינה כי אין למידע זה השפעה על דירוג המשתמשים. אני חושבת שהרוב המוחלט של המשתמשים לא מסתכל ולא מתייחס למידע זה לפני שהוא מוריד או מדרג אפליקציה.

בנוסף הורדתי את שתי העמודות המייצגות את מספר ההורדות ומספר הדירוגים של כל אפליקציה. אני חושבת שמידע זה היה יכול לשפר את חיזוי המודלים בצורה ניכרת, אך מכיוון שהפרויקט מנסה לחזות מה יהיה דירוג האפליקציות- זאת אומרת רמת ההצלחה שלהם, הוא בעיקרו מיועד לאנשים וחברות המעוניינים לפתח אפליקציות חדשות ורוצים לדעת לאיזה דברים כדאי להם לשים לב לפני שהם עושים זאת. לכן לקהל היעד אין יכולת לדעת מראש (לא בלי מחקר נוסף לפחות) מה יהיה מספר ההורדות ומספר הדירוגים.

שלב החיזוי:

השערת האפס שלי הייתה כי 60% מהאפליקציות ידורגו 4, ו 40% ידורגו 5. זאת מכיוון שהרוב המוחלט של האפליקציות בדאטה- 93% קיבלו דירוגים אלו (כאשר הדירוג 4 מהווה 58%, והדירוג 5 מהווה 35%) ושאר האפליקציות- האחוזים הנותרים דורגו 3, 2 או 1 (כאשר הדירוג 3 מהווה 5% מכלל הדאטה, והדירוגים 2 ו- 1 מהווים רק 2%).

המודלים שהשתמשתי בהם הם:

* KNN - אותו למדנו במהלך ההרצאות. לאחר שניסיתי מספר 'שכנים' שונה גיליתי כי מודל זה הגיע לרמת הדיוק הגבוהה ביותר כאשר מספר השכנים הוא 20. כמות גבוהה זאת של שכנים אפשרה לשיפור החיזוי בערכים הגבוהים יותר של הדירוג (4 ו- 5) מכיוון שבערכים אלו יש לנו את המספר הרב ביותר של אפליקציות, אך בערכים הנמוכים יותר לא הצליח להתבצע חיזוי כי כמות הערכים בדירוג זה שהיו ב test לא הספיקו. לכן רמת הדיוק הכללית של מודל זה לא עלתה בהרבה על רמת הדיוק כאשר מספר השכנים היה נמוך.
* Logistic Regression – מודל זה משתמש בפונקציה לוגריתמית המשתמשת במספר גורמים (הפיצ'רים) בכדי לבצע את החיזוי. מודל זה מתאים בעיקר לבעיות קלסיפיקציה בהן יש רק שני ערכים, אני מניחה שזאת הסיבה שבמודל זה החיזוי פעל על שני הערכים השכיחים ביותר, התוצאות היו כמעט זהות למודל הקודם.
* Random Forest – מודל זה משתמש במספר של עצי החלטה בכדי לבצע חיזוי. היתרון במספר רב של עצי החלטה הוא הגדלת ההתאמה של המודל לדאטה, ובכך ביצוע החיזוי משתפר, מבלי לגרום לעלייה ברמת ההטיה (bias) של המודל. מודל זה הגיע לתוצאות זהות למודלים הקודמים, אני מניחה שגם כאן החיזוי נפגע מכך שכמות הדירוג הנמוך בדאטה לא מאפשרת חיזוי מספק.