פרויקט ניהול מידע מבוזר – חלק ב׳ – Learning Task

רועי ברק ואלמוג בן סימון
206571135, 207375676

בהמשך לשאלה הראשונה, בה ביצענו ניתוחים ואנליזות על הנתונים שקיבלנו,
הסקנו כי ניתן לבצע מספר טרנספורמציות לנתונים כך שהמידע יהיה יותר אינפורמטיבי, ויותר פשוט ללמידה, כאשר נייצג את כל הפיצ׳רים כוקטור של מספרים.

תחילה ראינו כי העמודה של user, הינה בעלת ערכים המיוצגים כתווים. במשימת למידה כמובן שכל פיצ׳ר צריך להיות מיוצג באמצעות מספר, ולכן המרנו את ערכים אלו למספרים לפי ערכי האסקי של התווים.

מסקנה שנבעה לנו מהמשימה בה ביצענו אנליזות לנתונים, הינה כי ישנם מספר מודלים(models) בודדים (בדאטה שקיבלנו שמנו לב כי ישנו מודל יחיד, אך מכיוון שרצינו לבצע טרנספורמציה גנרית, עשינו זו באופן כללי כמספר בודד).
לכן המרנו עמודה זו למספר עמודות בינאריות כמספר המודלים השונים. כלומר, לכל מודל בנתונים ישנה עמודה כך לכל רשומה יהיה ערך 1 בעמודה אם היא בוצעה ממודל זה ואחרת 0.

מסקנה דומה הינה כי קיימים מספר מכשירים(devices) בודדים ולכן ביצענו טרנספורמציה דומה לטרנספורמציה שביצענו עבור הפיצ׳ר מודל(model).

כל שאר הפיצ׳רים היו בעלי ערכים אינפורמטיבים כגון מיקום לפי שלוש קורדינאטות(כל קורדינאטה בערך מספרי ולכן קל ללמוד איתה), זמני ההגעה והיצירה הינם בפורמט של unix time(number of seconds passed since 1970) ולכן מידע זה לעומת תאריכים שעה ושניות מאוד נוח מכיוון שמיוצג באמצעות מספר, ולכן אינפורמטיבי ללמידה.

לבסוף, שמנו לב כי הפיצ׳ר gt אשר הlabel של משימת הלמידה, הינו מיוצג באמצעות ערכים מסוג string. לכן המרנו ערכים אלו למספר שלם(חד חד ערכי).

לאחר הטרנספורמציות שביצענו על הנתונים, ניגשנו למשימת הלמידה. שקלנו מספר מודלים כך שיתאימו להתפלגות הנתונים שלנו, ובדקנו איזה מודל עם איזה פרמטרים הינו בעל התוצאות הטובות ביותר באמצעות Cross Validation. ניגשנו תחילה למודל Random Forest אשר נתן לנו ביצועים טובים מאוד, בטווח של 98-90 אחוז accuracy. לאחר מכן ניסינו מספר מודלים נוספים(Logistic Regression, Decision Tree, SVM) וקיבלנו תוצאות פחות טובות ולכן החלטנו להישאר עם מודל ה Random Forest.

לאחר בחירת המודל נשאר לנו לבחור את הפרמטרים הטובים ביותר שיגדירו את המודל שלנו. ביצענו cross validation(70 אחוז סט אימון, 30 אחוז ססט מבחן) עם מספר פרמטרים שונים, כגון עומק בטווח של 30-5, ומספר עצים בטווח של 20-5. כמובן שמטרתנו היא לבנות מודל מכליל אשר משיג דיוק גבוה מבלי להיות מורכב מידי ויגרום שגיאות זיכרון, ובנוסף להימנע מOverfitting, ולהימנע מעומק ומספר עצים גדול למען ייעול זמן ריצת הקוד.
לאחר הרצת Cross Validation בהתאם לפרמטרים, הסקנו כי המודל המתאים ביותר הינו Random Forest עם 5 עצים בעומק של 20.

לסיכום, במשימת הלמידה לאחר שלב האנליזות, הבנו אלו טרנספורמציות יעילות אנו צריכים לבצע על הנתונים כדי שמשימת הלמידה תהיה מדויקת ואפקטיבית ביותר. לאחר מכן ביצענו שלב ולידציה כדי לבחור את המודל המתאים ביותר לנתונים.

ולבסוף הגענו למודל החוזה על 30 אחוז(180 אלף רשומות) מהנתונים ברמת דיוק של 97-98%.