**פרויקט סיום – למידת מכונה**

חרות בן סלמון 208055277

ליעד הכלר 318382157

תאריך: 17/06/21

**תקציר מנהלים**

עבדנו בפרויקט עם דאטה המכיל נתונים על אנשים שהזמינו טיסות בין השנים 2015-2017.

המטרה הכללית שלנו הייתה לאמן מודל שמקבל את פרטי ההזמנה של נוסע מסוים וחוזה האם הנוסע צפוי לבטל או לא לבטל את הטיסה שלו.

תחילה קיבלנו 34 פיצ'רים, חלקם בעלי שם עם משמעות וחלקם אנונימיים. חקרנו את הדאטה והשקענו זמן רב ב-Pre-processing, על מנת לסדר ולעבד את הדאטה, תוך מחשבה על כך שעלינו להימנע מ-overfitting. במהלך התהליך, שינינו, הוספנו ומחקנו פיצ'רים, עד שהגענו ל-42 פיצ'רים סופיים מעובדים. מכאן, הורדנו את מספר הפיצ'רים ל-36, באמצעות בחירת פיצ'רים בעלי פוטנציאל לתרום יותר למודל.

לבסוף, ניסינו להריץ כמה מודלים שונים, וגילינו כי המודל שמסווג לנו בצורה הטובה ביותר הינו מודל Random Forest. קבענו זאת לפי מדד AUC, שיצא כ-10.92 עבור מודל זה.

דרך העבודה שלנו בהתחלה הייתה אינטואיטיבית ובכל שלב החלטנו על בסיס מה שהיה נראה לנו הכי הגיוני. בהמשך, חזרנו לכל נקודת החלטה וניסינו דרך אחרת, כך שבכל פעם ביצענו שינוי אחד. עבור כל דרך, בדקנו את תוצאות המודל וכך ידענו האם השינוי שעשינו שיפר לנו את המודל או לא. זה היה מעין מעגל שחזר על עצמו, עד שראינו שאין שיפור משמעותי בתוצאות המודל והגענו לנקודת העצירה שלנו.

**חלק ראשון - אקספלורציה**

* ערכים ריקים - הסתכלנו במבט כללי על כמות הערכים הריקים שיש בכל פיצ'ר, כדי להבין איזה פיצ'רים דורשים טיפול בערכים ריקים. (ראו נספח 1)
* מיון הפיצ'רים - חילקנו את הפיצ'רים לשלוש קטגוריות: פיצ'רים נומריים, פיצ'רים בולאניים ופיצ'רים קטגוריאליים. בשלב זה, לא החשבנו פיצ'רים בולאניים בעלי ערכים חסרים כפיצ'רים בולאניים, כיוון שעדיין לא ידענו איך נמלא את הערכים החסרים ואיך זה ישפיע על הקטגוריה של הפיצ'ר. מיון הפיצ'רים אפשר לנו בהמשך האקספלורציה לעשות בדיקות שונות עבור כל סוג פיצ'ר.
* התפלגות הערכים של פיצ'רים נומריים - החלטנו לבחון את ההתפלגויות של פיצ'רים נומריים, כדי להחליט איך כדאי לנו למלא תאים חסרים או האם כדאי לנו להשתמש בפיצ'רים אלו או לא. (ראו נספח 2)
* התפלגות הערכים של פיצ'רים בולאניים - רצינו לבדוק את ההתנהגות של פיצ'רים אלו כדי לדעת האם הם אינפורמטיביים או לא. כלומר, האם רוב הערכים של פיצ'ר מסוים זהים. (ראו נספח 3)
* קורלציה בין פיצ'רים נומריים - בדקנו את הקורלציה בין כל הפיצ'רים הנומריים. הנחנו כי זוג פיצ'רים בעלי קורלציה גבוהה מ-95% הינם פיצ'רים שמתנהגים מאוד דומה, ולכן מספיק לשמור רק אחד מהם. מצאנו כי ישנם שני פיצ'רים בעלי קורלציה גבוהה מה-threshold שבחרנו, ובחרנו למחוק את הפיצ'ר בעל קורלציה גבוהה יחסית עם פיצ'רים נוספים. (ראו נספח 4)
* חקירת פיצ'ר המדינות - בחנו כמה דוגמאות יש מכל מדינה. רצינו לבדוק מה טווח הדוגמאות שיש מכל מדינה וגילינו שקיימים פערים מאוד גדולים בין המדינות השונות. (ראו נספח 5)
* בחינת הקשר בין פיצ'ר ללייבל - יצרנו פונקציה שבודקת עבור כל אחד מהערכים של פיצ'ר מסוים כמה לייבלים יש מכל סוג. עשינו זאת כדי לבדוק את התפלגות הלייבלים עבור כל value בפיצ'ר, כדי להסיק מסקנות עבור דרך מילוי הערכים החסרים בשלב ה-preprocessing. (ראו נספח 6)
* מציאת "זנב" בהתפלגויות של פיצ'רים - יצרנו את הגרף במטרה לבדוק האם קיימים Outliers בפיצ'רים ספציפיים. מצאנו שעבור הפיצ'רים 'adults' ו-'anon\_feat\_1'- רוב הערכים מתרכזים בטווח מסוים, וקיים "זנב" להתפלגות. (ראו נספח 7)

**חלק שני - Preprocessing**

* פיצול הדאטה - תחילה פיצלנו את הדאטה לtrain ו-validation, כדי שנוכל לבצע בדיקות פנימיות עבור המודלים שנבחר להריץ.
  + החלטנו להשאיר 75% מהדאטה, רצינו שיהיו למודל מספיק דוגמאות להתאמן עליהן ועם זאת גם לtrain מספיק דוגמאות לבחון את המודלים שהתקבלו.
  + בחרנו לעשות את החלוקה בצורה רנדומלית כדי שנוכל לאמן את המודל ולבחון אותו על מדגם מייצג.
* מחיקת פיצ'רים לא אינפורמטיביים - הנחנו כי פיצ'ר עם למעלה מ-90% ערכים זהים אינו אינפורמטיבי למודל, ואף עשוי לגרום למודל לעשות Overfitting לדוגמאות האחרות.

ראינו לנכון למחוק את הפיצ'רים הבאים:

* + הפיצ'ר 'Unnamed: 0' - פיצ'ר זה משמש כ-ID עבור כל דגימה, ולכן אינו אינפורמטיבי למודל.
  + הפיצ'ר 'agent' - מלבד העובדה שהוא מכיל ערכים חסרים, הנחנו שאין משמעות למספרים בפיצ'ר זה, ערך עם מספר גדול יותר לא בהכרח טוב יותר ולא רצינו שהמודל יפגע מכך.
  + הפיצ'ר 'adr' - אנחנו חושבות כי פיצ'ר זה לא תורם לפרדיקציה של הלייבל, ולכן בחרנו למחוק אותו.
  + הפיצ'רים 'prev\_canceled' ,'prev\_not\_canceled', 'anon\_feat\_6' ו-'anon\_feat\_3' - בפיצ'רים אלה יש מעל 90% ערכים זהים, ובפרט שווים לאפס. כפי שציינו קודם, הנחנו כי פיצ'ר עם למעלה מ-90% ערכים זהים אינו אינפורמטיבי למודל, ואף עשוי לגרום למודל לעשות Overfitting לדוגמאות הנותרות.
  + פיצ'ר 'Company', 'anon\_feat\_13' - פיצ'רים עם אחוז גבוה של תאים ריקים אינם תורמים למודל. בפיצ'רים אלו יש למעלה מ-90% , הסף שהנחנו, ערכים ריקים ולכן בחרנו למחוק אותם.
* Outliers - פיצלנו את הטיפול ב-outliers לפיצ'רים נומריים ופיצ'רים קטגוריאליים:

פיצ'רים נומריים - בחרנו בשתי דרכי טיפול:

* + מחיקת outliers - ראינו שעבור הפיצ'רים  'adults' ו- 'anon\_feat\_1' , מרבית הערכים מתרכזים סביב התוחלת, בעוד קיימים Outliers. הנחנו שערכים שרחוקים יותר מ-3 סטיות תקן מהתוחלת הם ערכים קיצוניים ולכן מחקנו דוגמאות עם ערכים כאלו.
  + שינוי ערכים מרובים לערך יחיד - בהתאם לגרף 2 (ראו נספח 2), ראינו שקיים 'זנב' בהתפלגות הערכים של הפיצ'ר 'changes', החלטנו לאחד את הערכים בזנב לערך יחיד אותו בחרנו גם כן על פי גרף 2.

עבור פיצ'ר זה ראינו שרק ל-1.3% מהדוגמאות יש ערך שגבוה מ-3 ולכן החלטנו לבצע השמה של הערך זה.

פיצ'רים קטגוריאליים - גם כאן בחרנו לבצע זאת בשתי השיטות שציינו לעיל.

* + שינוי ערכים של outliers - עבור הפיצ'רים 'order\_type' ו- 'acquisition\_channel' מצאנו כי קיימות רק 4 דוגמאות בהן מופיע הערך 'Undefined' ולכן שינינו ערך זה לערך הכי שכיח בכל פיצ'ר.
  + שינוי ערכים מרובים לערך יחיד - ביצענו זאת עבור הפיצ'ר 'country' . החלטנו לקבוע שעבור מדינות שבהן מספר הדוגמאות הוא נמוך מ-1% ממספר הדוגמאות הכולל, נשנה את הערך שלהן ל-'Other'. כך, כאשר אנחנו יוצרות dummy variables לא נקבל מספר מאוד גבוה של פיצ'רים בגלל שיש הרבה ערכים שונים בפיצ'ר המדינות.

עבור הדאטה של ה-test- גילינו שיש מדינות שנמצאות בדאטה זה ולא נמצאות בדאטה של ה-Train. מדובר במספר מאוד קטן של דוגמאות ממדינות אלה, לכן שינינו את ערכן גם כן ל-‘other’.

* שינוי ה-type של פיצ'רים -
  + עבור הפיצ'ר 'order\_week' - הפכנו את הערכים מ-string ל-int.
  + עבור הפיצ'ר 'anon\_feat\_12' - שינינו את הערך True ל-1, ואת הערך False ל-0, כי כשאנחנו עושות dummy variables, לא נרצה שהפיצ'ר יהפוך לשני פיצ'רים.
  + עבור הפיצ'ר 'order\_month' - הפכנו את החודשים מ-string ל-int.
* מילוי ערכים ריקים -

עבור פיצ'רים נומריים ובולאניים:

עבור חלק מהפיצ'רים ניסינו שתי שיטות- למלא לפי הרוב ולמלא לפי הממוצע.

* + עבור חלק מהפיצ'רים ראינו שערכיהם יחסית מפוזרים ולכן מילאנו את הערכים החסרים לפי הממוצע.
  + עבור פיצ'רים אחרים ראינו שהרוב המוחץ של הערכים שווה לערך מסוים, ולכן ניסינו למלא ערכים ריקים לפי הרוב, כיוון שלא רצינו לשנות את התפלגות הערכים של הפיצ'ר.

לאחר ניסיונות, גילינו שמילוי ערכים ריקים בממוצע נתן לנו מודל טוב יותר. לכן, שינינו מהרוב לממוצע.

עבור פיצ'רים קטגוריאליים:

* + עבור הפיצ'ר 'order\_month':

ניסינו למלא את הערכים הריקים בשתי שיטות:

* + - ראינו שהיחס בין הלייבלים של תאים לא ריקים עבור כל חודש הוא יחסית זהה (כ-66% False). בנוסף, חישבנו את מספר הלייבלים שהם False בתאים הריקים וגילינו שהיחס הוא כמעט זהה. לכן, ניסינו למלא את התאים הריקים בערכים 1-12, כך שהחלוקה תהיה יחסית שווה בין הלייבל False לבין הלייבל True בכל חודש כדי לשמור על היחס.
    - נעזרנו בפיצ'ר 'order\_week' כדי למלא את הערכים החסרים בפיצ'ר 'order\_month'. עבור דגימה שמספר השבוע שלה בין 1-4 מילאנו 1, עבור דגימה שהשבוע שלה בין 4-8 מילאנו 2, וכן הלאה.

לאחר ניסיונות הרצה, גילינו שהשיטה השנייה עובדת טוב יותר ולכן בחרנו בה.

בנוסף, כשהגענו לשלב החלת ה-Preprocessing על הטסט, הבנו שלא נוכל להשתמש בשיטה הראשונה כיוון שהיא מסתמכת על כך שיש לנו את ה-labels.

* + עבור הפיצ'רים 'Deposit\_type' ו-'customer\_type' - יצרנו ערך חדש 'other'. ראינו שעבור כל ערך בפיצ'רים האלה, הלייבלים מתנהגים בצורה דומה (ראו נספח 6). לא רצינו לשנות את ההתנהגות של הפיצ'רים ולכן החלטנו ליצור ערך חדש עבור התאים הריקים.
* יצירת פיצ'רים חדשים -
  + החלטנו לאחד את הפיצ'רים 'adults', 'babies' ו-'children' לפיצ'ר אחד שהוא סכום של שלושתם, וקראנו לפיצ'ר החדש 'travelers'.

עבור הדאטה של ה-train, מחקנו דוגמאות שבהן הערך של ‘travelers’ הוא 0, כיוון שמבחינתנו הן outliers. נציין כי מדובר במספר מאוד קטן של דוגמאות.

* + שינינו את הערכים של הפיצ'ר 'time\_until\_order' מימים למספרי חודשים.
  + שינינו את הערכים של הפיצ'רים הקטגוריאליים בעלי 5 ומטה ערכים שונים באמצעות OrdinalEncoder- שמשנה ערכים קטגוריאליים לערכים מספריים בטווח של 0 עד מספר הקטגוריות שבפיצ'ר פחות אחד.
* יצירת dummy variables - הפכנו את כל הפיצ'רים לפיצ'רים נומריים.
* כדי לוודא את עצמנו- יצרנו גרף boxplot כדי לבדוק שאחרי תהליך ה-preprocessing אכן נשארנו עם מספר מאוד נמוך של outliers. בנוסף, רצינו לוודא שאין לנו ערכים מאוד קיצוניים שעלולים לפגוע ב- MinMax Scaling, שכן הוא רגיש ל-outliers בשל הסתמכותו על הערך המקסימלי והערך המינימאלי. (ראו נספח 8)
* Scaling - ראינו שטווח הערכים של כל פיצ'ר שונה מאוד מטווח הערכים של שאר הפיצ'רים. ללא נרמול, הפיצ'רים לא יתרמו בצורה שווה למודל, וכן נרמול מסייע לחלק מהמודלים לרוץ מהר יותר.

ניסינו שתי דרכים לנרמול הדאטה:

* + סטנדרטיזציה- קיבלנו טווח קטן של ערכים עם תוחלת אפס ושונות ששווה לאחד.
  + MinMax Standardization - קיבלנו טווח אחיד של כל הערכים בדאטה.

לאחר ניסיונות הרצה, גילינו שעם MinMax Standardization קיבלנו מודלים טובים יותר, ולכן בחרנו לנרמל בדרך זו.

* Feature Selection -

הסיבות להורדת המימדיות- נרצה להימנע מ-overfitting על הדאטה הספציפי שלנו, כדי שהמודל ידע להכליל טוב יותר על דוגמאות חדשות. נרצה שה-AUC של ה-train יהיה קרוב ל-AUC של ה-validation, וכך אכן יצא לנו. כלומר, התאמנו על דאטה מסוים, ונתנו למודל דוגמאות חדשות והוא ידע לחזות עליהן טוב.

על מנת להגדיל את יכולת ההכללה של המודל שלנו, השתדלנו להוריד את כמות הפיצ'רים בצורה הגיונית ולשמור פיצ'רים בעלי קורלציה גבוהה ללייבלים.

טכניקות שבדקנו להורדת המימדיות של הבעיה/Feature Selection:

* + PCA- ניסינו כמה ניסיונות של הפחתת מימדים באמצעות טכניקת PCA. אולם, ללא PCA המודל שלנו נתן ביצועים טובים יותר. לכן, בחרנו לעשות Feature Selection בצורה אחרת.
  + KBest- דרך לבחירת K הפיצ'רים הטובים ביותר, לפי מבחן חי בריבוע. בחרנו בדרך זו, כיוון שהיא נתנה לנו ביצועים טובים.

**חלק שלישי - הרצת מודלים**

* מודלים ראשוניים - בחרנו להריץ את המודלים:
  + KNN
  + Logistic Regression
* מודלים מתקדמים - בחרנו להריץ את המודלים:
  + Random Forest
  + Multi-Layer Perceptron

**חלק רביעי - הערכת המודלים**

* הערכת המודל באמצעות K-Fold Cross Validation - הערכנו כל אחד מהמודלים באמצעות דרך זו, השתמשנו ב-k=10 כדי לעמוד בזמן הריצה הכולל של הפרויקט.

הצגנו בגרף עבור כל מודל את כל אחד מה-roc curves שיצא עבור כל K-fold וגם הראנו את ה-roc curve הממוצע.

עבור כל מודל, ה-roc curves השונים של כל K-fold יצאו לנו קרובים, מה שעשוי לרמז על מודל יציב.

* בניית Confusion matrix על אחד המודלים - בחרנו להראות את ה-confusion matrix על המודל Random Forest:
  + TP - מספר הדוגמאות שהלייבל שלהן הוא TRUE ושהגיעו לעלה שמסווג כ-TRUE.
  + TN - מספר הדוגמאות שהלייבל שלהן הוא FALSE ושהגיעו לעלה שמסווג כ-FALSE.
  + FP - מספר הדוגמאות שהלייבל שלהן הוא FALSE אבל הן הגיעו לעלה שמסווג כ-TRUE.
  + FN - מספר הדוגמאות שהלייבל שלהן הוא TRUE אבל הן הגיעו לעלה שמסווג כ-FALSE.
* פערי ביצועים בין הרצת הומדל על ה-train לבין הרצת המודל על ה-validation - ניסינו להימנע מ-overfitting, כדי שהמודל שלנו ידע להכליל טוב יותר על דוגמאות חדשות. נרצה שה-AUC של ה-train יהיה קרוב ל-AUC של ה-validation, וזה אכן מה שקיבלנו. כלומר, המודל התאמן על דאטה מסוים וכשהוא קיבל דוגמאות חדשות הוא סיווג אותן בצורה טובה.

על מנת להגדיל את יכולת ההכללה של המודל, ניסינו להוריד את כמות הפיצ'רים בצורה הגיונית ולשמור על פיצ'רים בעלי קורלציה גבוהה עם הלייבלים.

**סיכום ומסקנות**

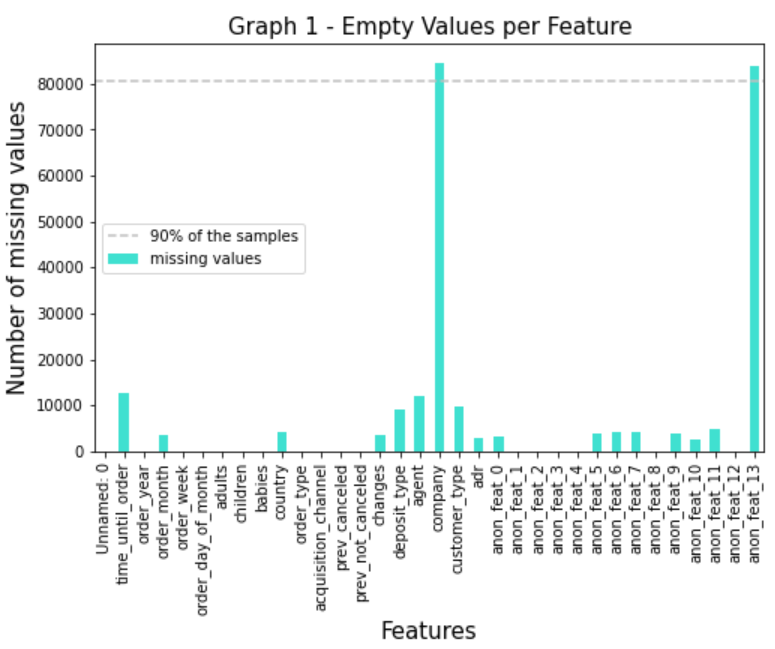
לסיכום, לאחר ניסיונות שונים של הרצה, למדנו להכיר את הדאטה שקיבלנו וכן למדנו אילו שינויים בכל שלב ב-Preprocessing שיפרו את המודל שבחרנו ואילו מתוכם היו הכי משמעותיים.

המודל שבחרנו הוא Random Forest אשר נתן לנו את הביצועים הטובים ביותר על הדאטה שלנו.

מסקנות מרכזיות:

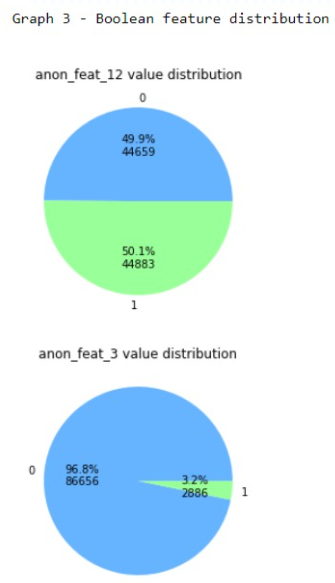
* גילינו כי הרצת המודלים עם מספר קטן של פיצ'רים הובילה לתוצאות פחות טובות. מכך, הסקנו כי עבור הדאטה הזה, יש צורך במספיק פיצ'רים על מנת להגיע להרבה פרדיקציות נכונות. הרצת מודלים עם פחות מ-10 פיצ'רים הניבו תוצאות לא טובות על ה-validation.
* הסקנו כי המודלים הפשוטים- רגרסיה לוגיסטית ו-KNN, עובדים פחות טוב עבור הדאטה הזה, כיוון שמדובר בדאטה מורכב עם הרבה פיצ'רים שהם שונים זה מזה. לכן, מודל פשוט לא ידע להפריד בצורה הכי טובה בין אנשים שביטלו טיסה לבין כאלה שלא, בשל מורכבות הבעיה.

נספחים

**נספח 1** – גרף המייצג את מספר הערכים החסרים בכל פיצ'ר

A picture containing chart

Description automatically generated**נספח 2** – גרף המייצג את התפלגות הפיצ'רים הנומריים

**נספח 3** – התפלגות פיצ'רים בולאניים

Chart

Description automatically generated**נספח 4** – מפת קורלציה בין פיצ'רים נומריים

Chart, bar chart, histogram

Description automatically generated**נספח 5** – התפלגות מספר הדוגמאות מכל מדינה

Chart, bar chart

Description automatically generatedChart, bar chart

Description automatically generated**נספח 6** – לוג התפלגות הלייבלים עבור חלק מהפיצ'רים

**נספח 7** – התפלגות של פיצ'רים עם "זנב"

Chart, histogram

Description automatically generated**נספח 8** – גרף להצגת ה-Outliers

Chart, waterfall chart, box and whisker chart

Description automatically generated