מבוא ללמידת מכונה
מדעים דיגיטליים להייטק
אוניברסיטת תל אביב
אביב 2020-2021
מרצה: דור בנק
מתרגל: שחף גורן

נספחים:

נספח א': התפלגויות פיצ'רים ביחס לליבלים
נספח ב': גרפים של Acquisition Channel ו-Order Type
נספח ג': גרף המדינות המופיעות ביותר מ-1% מהדאטה
נספח ד': גרף AUC של מודל Random Forest על סט ה-Validation
נספח ה': גרפי Boxplot של פיצ'רים בעלי Outliers
נספח ו': confusion matrix למודל logistic regression על סט ה-validation
נספח ז': Grid Search

דו"ח פרויקט מסכם - זוג 16

פרטי המגישים:
נועה שפירא, 319121166
תומר בלום, 208947382

תקציר מנהלים:

א. בפרויקט זה, עלינו לפתור בעיה בתחום התיירות. קיבלנו נתונים אודות הזמנות של עסק בתחום התיירות. בענף זה ישנם ביטולים רבים המקשים על ניהול ההזמנות, תפקידנו לבנות מערכת **המנבאת את סיכוייה של הזמנה עתידית להתבטל**.

ב. קיבלנו דאטה בתצורת קובץ CSV המכיל כ-89,000 הזמנות ועבור כל הזמנה מפרט מידע כמו תאריך ההזמנה, הרכב הנוסעים, מדינה, מידע אודות ביטולים קודמים וכן פיצ'רים אנונימיים איתם **נאלצנו להתמודד בעזרת כלים שונים** כמו ציור גרפים וחישוב קורלציות.

ג. לאחר הרצת כלל המודלים האופציונליים, בחרנו במודל **Random Forest**, מכיוון שאיתו השגנו את **התוצאות הגבוהות ביותר**, בזמן ריצה קצר יחסית. במודל זה קיבלנו AUC של 0.93 עבור פרדיקציות על סט ה-Validation (נספח ד')

ד. במהלך עבודתנו, ניסינו שיטות שונות לשפר את התוצאות מצד אחד ולייעל את זמן הריצה מצד שני. לדוגמא, את עמודת המדינות שינינו כך שכל מדינה שמופיעה בפחות מ-1% בסט ה-Train תרשם בתור 'other'. הצלחנו לצמצם את מספר המדינות מכ-160 ל-14 וכתוצאה מכך מספר הקטגוריות בפיצ'ר ירד דרסטית, ולאחר מכן כשהשתמשנו ב-one hot encoder, מס' המימדים ירד גם כן. בנוסף, יצרנו משתנה חדש הסוכם את כלל האנשים בהזמנה (num_of_people) ופיצ'ר בינארי הבודק אם אנשים ישלמו על ביטול (deposit_given).

ה. במהלך הפרויקט שינינו קטעי קוד רבים משיקולי זמן ריצה והגענו לזמן ריצה סופי של כ-28 דקות.

ו. לסיכום, במהלך הפרויקט השתמשנו בטכניקות שנלמדו בהרצאות ובתרגולים וכן בשיטות שונות שלמדנו באינטרנט. למדנו מהעשייה וגילינו שיטות יצירתיות לעיבוד הדאטה.

מהלך הפרויקט

שלב א' - אקספלורציה:

הנחות שנלקחו בשלב זה:

- בשלב זה השתמשנו בכל הדאטה (כלומר, ללא פיצול ל-Train ו-Validation), מכיוון שכמות הדאטה מוגבלת ורצינו לקבל תמונה ברורה של התפלגויות הפיצ'רים והקורלציות ביניהם.
- מהתבוננות בהתפלגויות הפיצ'רים השונים, רואים כי ל-time_until_order, anon_feat_6, adr, adults, children ו-babies יש outliers. בהמשך נשקול את אופי הטיפול בהם.

- בשלב זה שינינו חלק מהפיצ'רים באופן כזה שיאפשר לנו להציג אותם בצורה גרפית: את הפיצ'ר 'order_month' עדכנו כך שהחודשים יופיעו כמספרים ואת הפיצ'ר 'order_week' עדכנו כך שמחקנו את החלק השמאלי ('week_') והפכנו ל-int.
לדעתנו, שינוי בשלב כזה (מחוץ לעיבוד המקדים) הוא תקין מכיוון שהחודשים ידועים מראש ולכן המיפוי יכול להיעשות בצורה זהה על סט ה-Test ובשבועות ביצענו מחיקה של המלל ולא של מספר השבוע.

<u>פירוט שלבי האקספלורציה</u>:

- את שלב האקספלורציה חילקנו לשני תתי שלבים, בשלב הראשון ניתחנו את הדאטה ללא הלייבלים ובשלב השני הצגנו את ההתפלגות של הפיצ'רים ביחס ללייבלים.
- ראשית, קראנו את קובץ feature_data וראינו מאילו פיצ'רים הוא מורכב.
- בדקנו מה הם סוגי המשתנים של כל פיצ'ר כדי שנוכל להתייחס לכל אחד בהתאם לסוגו.
- המשכנו בלחלק את הפיצ'רים לשתי קבוצות, מספריים וקטגוריאליים. הפיצ'ר של מספר ההזמנה הוא בעל שונות גבוהה ולכן לטעמנו אין צורך להשתמש בו במודל, נצטרך אותו בהמשך על מנת לכתוב את קובץ הפרדיקציות הסופי ולכן הורדנו אותו מהקבוצה של המשתנים המספריים.
- ציירנו היסטוגרמות של כל אחד מהפיצ'רים המספריים על מנת להבין את ההתפלגות של כל אחד מהם בצורה מיטבית.
- על מנת שנוכל להתמודד עם ערכים חסרים בדאטה נבדוק כמה ערכים חסרים יש בכל פיצ'ר. שמנו לב כי ל-company ול-anon_feat_13 יש מעל ל-80,000 ערכים חסרים, כמעט כאורך הדאטה, לכן נשקול להורידם בשלב העיבוד המקדים.
- בדקנו כמה מדינות שונות מופיעות בדאטה (163), מכיוון שהשונות של פיצ'ר זה היא גבוהה נצטרך לשקול להורידו / לשנותו. הרעיון שעולה לנו כרגע, הוא לקודד לקטגוריה אחת מדינות שמופיעות בפחות מאחוז מהדאטה. ביצענו בדיקה והצגה של המדינות שמופיעות ביותר מאחוז, אותן נרצה להשאיר (נספח ג').
- לדעתנו, הפיצ'רים order_week ו-order_month יהיו בעלי קורלציה גבוהה, לכן ציירנו כל אחד מהם ביחס למספר ההזמנות ושמנו לב שקיים דמיון רב בין הגרפים וכי השיא מגיע בערך באותה נקודה. הנחנו שקיימים חודש ושבוע בהם מספר ההזמנות הוא הגבוה ביותר, בהמשך נבדוק אם קיים קשר בין הלייבל לבין התפלגות כל פיצ'ר.
- בדקנו את הקשר בין המשתנים בעזרת מטריצת קורלציות, מצאנו שיש קשר בין order_week ל-order_month, בין time_until_order ל-anon_feat_11 ובין agent ל-anon_feat_9. ציירנו scatter plot לכל אחד משלוש הזוגות, הגרף המעניין ביותר הוא הגרף שמייצג את הקשר בין time_until_order ל-anon_feat_11, ניתן לראות כי קיים מעין קשר לינארי.
- ציירנו את הפיצ'רים הקטגוריאליים הנותרים והגענו לשלוש מסקנות: סוג הלקוח הנפוץ ביותר הוא לקוח ארעי, הערכים בפיצ'ר acquisition_channel דומים ביותר לערכים של פיצ'ר order_type ונשקול להוריד את אחד מהם (נספח ב') וב-anon_feat_7 היחס בין 0 ל-1 גבוה ביותר, נשקול להסיר את הפיצ'ר בהמשך.
- ציירנו boxplot לכל הפיצ'רים מההנחות בתחילת הסעיף, על מנת להחליט אילו outliers נרצה להוריד (נספח ה').
- בשביל להבין אילו פיצ'רים יעזרו לנו לסווג יותר ואילו פחות, ציירנו אותם ביחס ללייבלים (נספח א'). במבט ראשוני, ניתן לראות כי הפיצ'רים order_month, order_year, order_week, order_day_of_month, agent, anon_feat_2, anon_feat_3 ו-anon_feat_6 ככל הנראה לא יועילו לנו בסיווג.

שלב ב' - עיבוד מקדים:

הנחות שנלקחו בשלב זה:

- בשלב זה , הנחנו כי עמודת מספרי ההזמנות אינה תעזור לביצוע פרדיקציות, מכיוון שהיא בעלת שונות גבוהה מאוד ולכן הסרנו אותה.
- בהתאם לגרפים בשלב האקספלורציה (נספח א'), החלטנו להסיר את פיצ'ר agent. לפני כן, ניסינו לקודד אותו בצורה דומה לקידוד שביצענו בפיצ'ר country ובגלל שלא הייתה הטבה ב-AUC החלטנו להורידו.
- בגרפים בשלב האקספלורציה (נספח א') ראינו כי anon_feat_2, anon_feat_3, anon_feat_6, order_month, order_day_of_month, מתפלגים בצורה שאינה תורמת להבחנה בין הלייבלים ולכן החלטנו להורידם (לשם בדיקה, ביצענו הרצות עם הפיצ'רים ובלעדיהם – וה-AUC נשאר זהה).
- בהמשך לשלב האקספלורציה, שם ראינו בגרפים של הפיצ'רים acquisition_channel ו-order_type (נספח ב') כי ערכי הפיצ'רים דומים, החלטנו להסיר את הפיצ'ר acquisition_channel. בחרנו בפיצ'ר זה מבין השניים מכיוון שהוא מכיל פחות סוגי ערכים מ-order_type.
- בחרנו להוריד outliers אחרי פיצול הדאטה ומסט ה-train בלבד, מכיוון שמה-test לא נוריד outliers ורצינו שה-validation ישקף בצורה טובה את ה-auc האמיתי.
- שמנו לב שבכ-100 הזמנות סך האנשים בהזמנה הוא 0 (0 מבוגרים, 0 ילדים ו-0 תינוקות), דבר שנראה לנו מעט לא הגיוני ובחרנו להסיר שורות אלה (אך ורק מסט ה-train).

החלפנו את הלייבלים מ-True ו-False ל-1 ו-0 (בהתאמה) והגדרנו מספר פונקציות שיעזרו לנו לעבד את הדאטה:

- פונקציה שמורידה outliers.
- מספר פונקציות למילוי ערכים חסרים (לפי חציון, ממוצע, השמת 0 והשמת מילה).
- פונקציה (ופונקציית עזר) לקידוד מחדש של פיצ'ר country (מקודד את כל המדינות שמופיעות בפחות מ-1% מה-Train / לא מופיעות בו כלל, ל-other).
- פונקציה שמחליפה בפיצ'ר deposit_type ל-1 עבור non Refund ו-0 אחרת.
- קידוד משתנים קטגוריאליים באמצעות ()one_hot_encoder.
- סטנדרטיזציה באמצעות ()standard_scaler.
- הפחתת ממדים באמצעות ()pca.

פיצלנו את הדאטה לסט Train (80%) וסט Validation (20%).

מהלך העיבוד המקדים:

<u>מחיקת Outliers</u>:

בהתאם לגרפי ה-boxplot (נספח ה') החלטנו להשאיר outliers של הפיצ'ר time_until_order כדי לא למחוק אחוז גדול מהדאטה סט. בפיצ'ר adr מחקנו outliers לפי 3 סטיות תקן ובפיצ'רים children, adults ו-babies מחקנו שורות שלפחות אחד הערכים בהן גדול או שווה ל-10. בהמשך העיבוד המקדים, לפני ה-pca, ביצענו נרמול של הדאטה. נרמול שכזה אמור להפחית את השפעת ה-outliers שלא מחקנו.

בחרנו לבצע את העיבוד המקדים על סט ה-Train בנפרד בגלל שצריך לבצע את העיבוד המקדים בהתאם לסט ה-Train. (תחילה ביצענו זאת בפונקציה אחת בעזרת משתנה אינדיקטור, אך הקוד היה מסורבל ולא קריא).

העיבוד המקדים מחולק למקטע קוד של סט ה-Train ופונקציה כמעט זהה עבור הסטים האחרים.

- מילוי הפיצ'רים: מס' ילדים, שינויים, חודש הזמנה ופיצ'ר אנונימי 7 ב-0. עבור 2 הפיצ'רים הראשונים, הנחנו שאם לא מפורט ערך כלשהו, הוא ככל הנראה 0, עבור חודש ההזמנה קידדנו 0 עבור חודש לא ידוע ועבור פיצ'ר אנונימי 7, ראינו כי הוא בינארי וכי הרוב המוחלט של הערכים הם 0, ולכן החלטנו לקודד כ-0 כדי למנוע מניפולציה שגויה של הדאטה.
- מילוי הפיצ'רים: זמן עד הזמנה, adr והפיצ'רים האנונימיים 0/5/9/10/11 בערך הממוצע של הפיצ'ר. הערכים האלו אינם מספרים שלמים ולכן מילוי הממוצע הוא גם לוגי וגם אינו משנה את הממוצע הקיים.
- מילוי פיצ'רים קטגוריאליים: עבור הפיצ'רים הקטגוריאליים החלטנו למלא ערכים חסרים במילה other ובכך ליצור קטגוריה חדשה לערכים החסרים. עבור פיצ'ר המדינות החלטנו להחליף ערכים חסרים במדינה פורטוגל (הנפוצה ביותר) כדי לא להשפיע על האופן בו נרצה לקודד את המדינות (לפי אחוזים מהדאטה).
- מחקנו 4 פיצ'רים: anon_feat_13 ו-company שהיו עם יותר מ-85,000 ערכים חסרים, עמודת מספרי ההזמנה שהיא בעלת שונות גבוהה מאוד וככל הנראה לא תעזור לנו עם הפרדיקציות ופיצ'ר agent, שבשלב הניתוח בהשוואה ללייבלים ראינו כי אין הפרדה כלל הפרדה ל-True ו-False וכי הוא ככל הנראה לא יועיל. בנוסף, מחקנו את הפיצ'רים שרשמנו בהנחות העיבוד המקדים.
- יצרנו פיצ'ר חדש בשם num_of_people, שמייצג את סה"כ האנשים בהזמנה, ומחקנו את השורות בהן ערך זה היה 0 (סט ה-train בלבד).
- יצרנו פיצ'ר חדש בשם deposit_type, שמייצג אם הזמנה היא ללא החזר.
- בדומה לקידודים קודמים, עדכנו את anon_feat_12 הבוליאני ל-1 ו-0.
- **בשלב זה, בעיבוד סט ה-Train בלבד, יצרנו עותק של X_train על מנת שנוכל להשתמש בו בהמשך, לעיבוד המקדים של סטי ה-Validation וה-Test.**
- קידוד הפיצ'רים הקטגוריאליים באמצעות one hot encoder.
- סטנדרטיזציה של הדאטה.
- כעת ביצענו הפחתת ממדים באמצעות PCA (בפונקציה נפרדת מטעמי נוחות), בחרנו ב-PCA מכיוון שראינו שיש קורלציה בין פיצ'רים שעלולה להפריע לפרדיקציות. בחרנו לשמור על 99% מהשונות המוסברת על מנת לקבל תוצאות טובות יותר, ועבורה קיבלנו הפחתה ל-31 ממדים.

שלב ג' - הרצת מודלים:

בשלב זה, הרצנו את כל המודלים האפשריים ולאחר בחינת זמן הריצה של כל מודל ותוצאת ה-AUC, בחרנו את ארבעת המודלים הבאים:

<u>מודלים ראשוניים</u>:

- Logistic Regression
- K-Nearest Neighbors

<u>מודלים מתקדמים</u>:

- Random Forest
- Multi-Layer Perception (ANN)

כדי למצוא את ההיפר פרמטרים האידיאליים עבור כל מודל ביצענו בדיקה בעזרת Grid Search, את הקוד השארנו כהערה בסוף שלב ד' והתוצאה מודפסת במחברת (ובנספח ז').

שלב ד' - הערכת מודלים:

ראשית, ציירנו confusion matrix למודל Logistic Regression על סט ה-Validation (נספח ו') ופירטנו על התאים.
לאחר מכן, הגדרנו והרצנו פונקציית K-Fold Cross Validation על כל אחד מהמודלים שבחרנו. את ה-AUC הגבוה ביותר קיבלנו עבור מודל Random Forest (0.927).
לבסוף, בדקנו פערי ביצועים בין סט ה-Train לסט ה-Validation עבור כל אחד מהמודלים, כאן את ה-AUC הגבוה ביותר לסט ה-Validation קיבלנו עבור מודל Random Forest ו-MLP (0.93).

שלב ה' - ביצוע פרדיקציה:

קראנו את קובץ הטקסט וביצענו עליו עיבוד מקדים, כעת נחבר את סט ה-Train וסט ה-Validation (וגם את הלייבלים) כדי שהמודל יוכל להתאמן על כמה שיותר מידע. ניעזר בפונקציית predict_proba כדי למצוא את הפרדיקציות ונכתוב לקובץ csv בפורמט המתאים.

<u>סיכום</u>

נסקור את המודלים אותם בחרנו:

K-NN:

- במודל זה קיבלנו ROC ממוצע של 0.869 בשיטת K-Fold Cross Validation.
- בביצוע פרדיקציות על סט Train וסט Validation, קיבלנו AUC של 0.91 ב-Validation ו-AUC של 1.00 ב-Train.
- ההיפר פרמטרים שקיבלנו בעזרת GridSearch הם: {'n_neighbors': 50, 'weights': 'distance'}

Logistic Regression:

- במודל זה קיבלנו ROC ממוצע של 0.853 בשיטת K-Fold Cross Validation.
- בביצוע פרדיקציות על סט Train וסט Validation, קיבלנו AUC של 0.89 ב-Validation ו-AUC של 0.85 ב-Train.
- ההיפר פרמטרים שקיבלנו בעזרת GridSearch הם: {'C': 1000, 'penalty': 'l2'}

Random Forest:

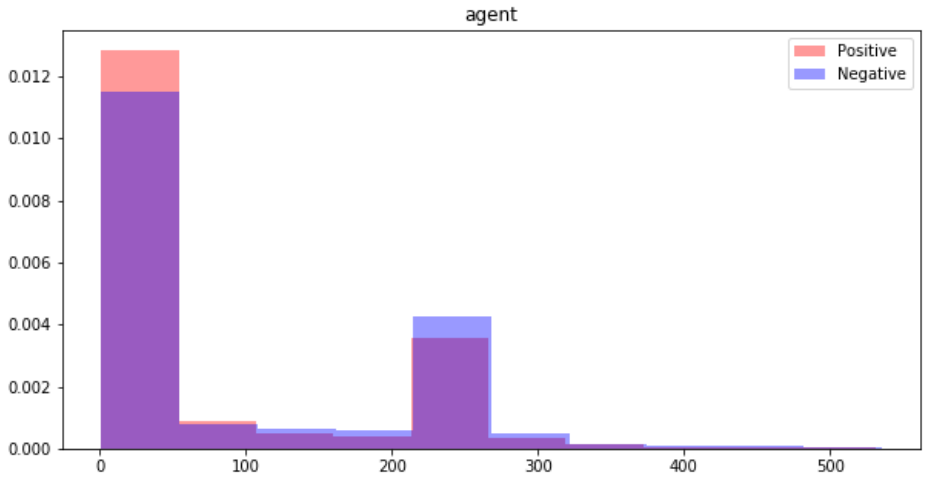
- במודל זה קיבלנו ROC ממוצע של 0.887 בשיטת K-Fold Cross Validation.
- בביצוע פרדיקציות על סט Train וסט Validation, קיבלנו AUC של 0.93 ב-Validation ו-AUC של 1.00 ב-Train.
- ההיפר פרמטרים שקיבלנו בעזרת GridSearch הם:
{criterion': 'gini', 'min_samples_leaf': 3, 'min_samples_split': 2, 'n_estimators': 200'}

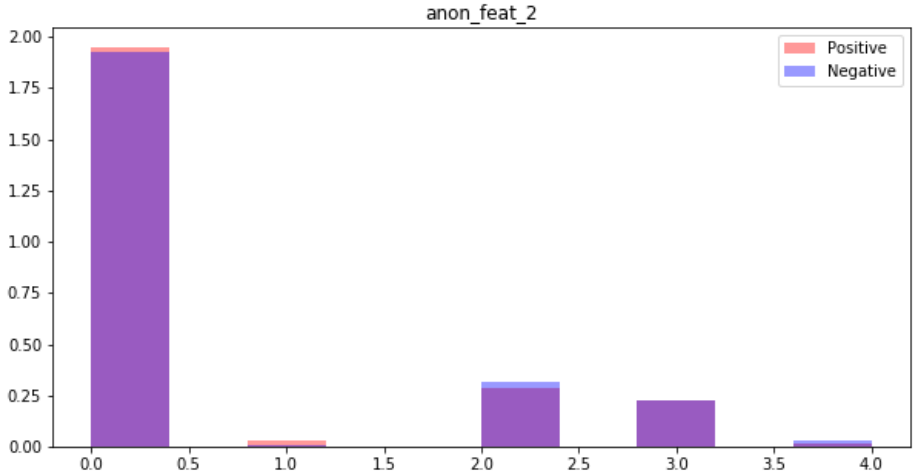
MLP:

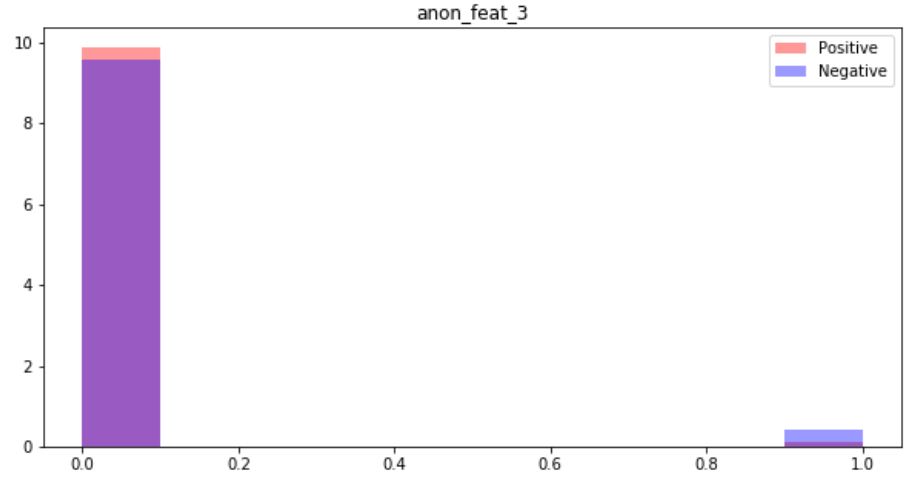
- במודל זה קיבלנו ROC ממוצע של 0.887 בשיטת K-Fold Cross Validation.
- בביצוע פרדיקציות על סט Train וסט Validation, קיבלנו AUC של 0.92 ב-Validation ו-AUC של 0.90 ב-Train.
- ההיפר פרמטרים שקיבלנו בעזרת GridSearch הם:
{activation': 'logistic', 'batch_size': 10, 'hidden_layer_sizes': (50, 50), 'learning_rate_init': 0.01, 'max_iter': 1500}

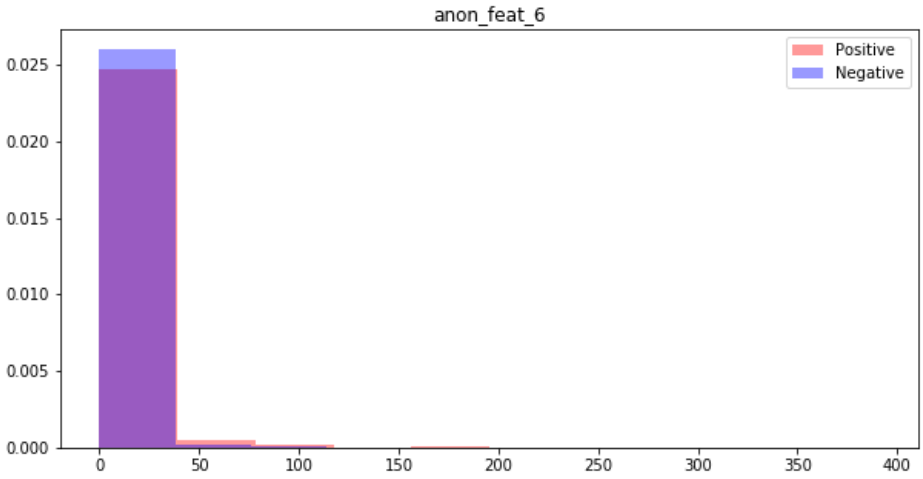
מבוא ללמידת מכונה
מדעים דיגיטליים להייטק
אוניברסיטת תל אביב
אביב 2020-2021
מרצה: דור בנק
מתרגל: שחף גורן

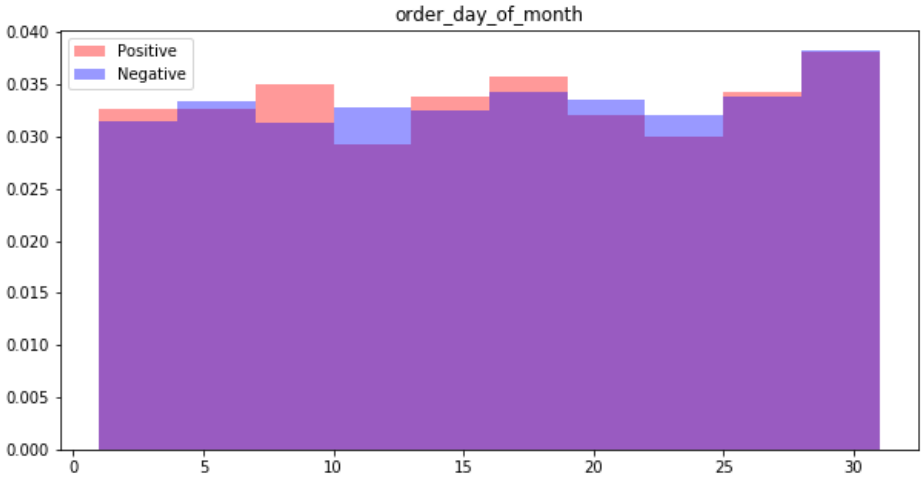
נספח א': התפלגויות פיצ'רים ביחס לליבלים

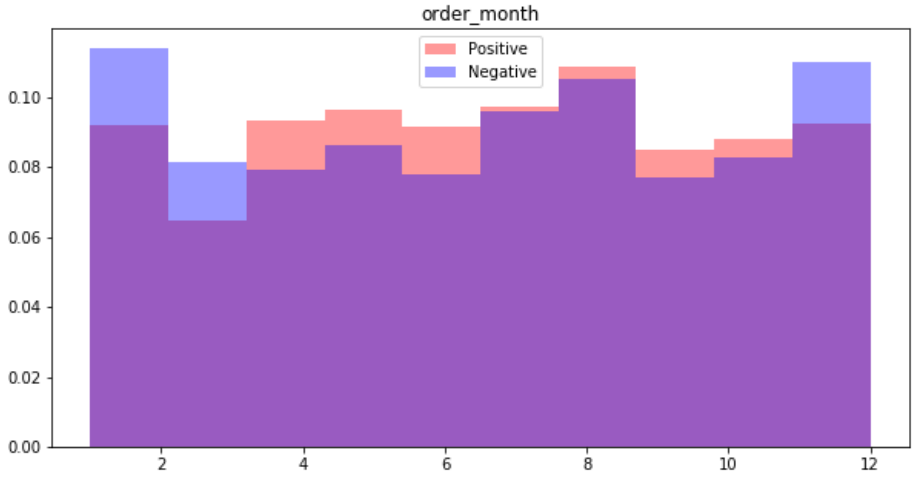




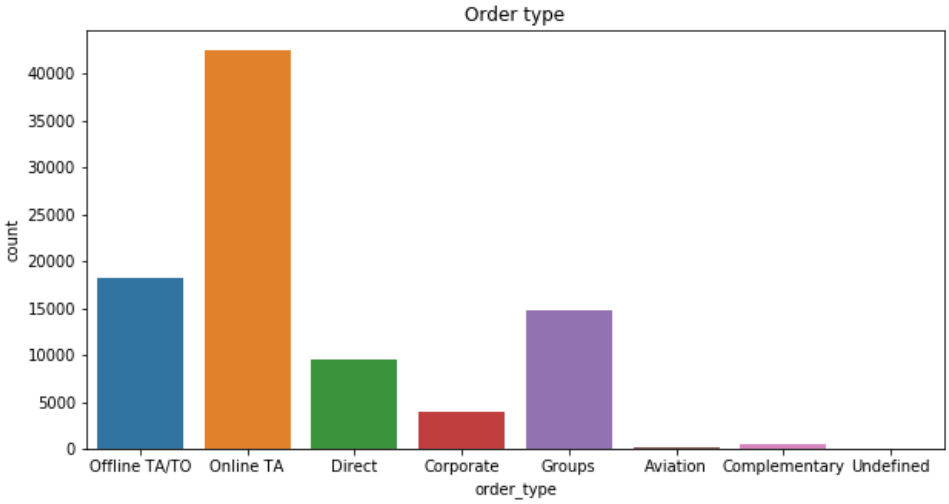


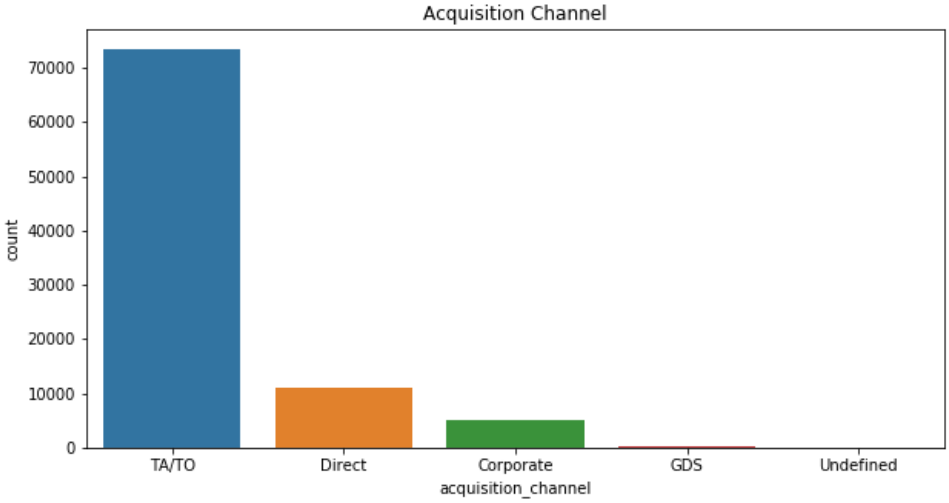

anon_feat_6
Positive
Negative
0.025
0.020
0.015
0.010
0.005
0.000
0
50
100
150
200
250
300
350
400


order_day_of_month
Positive
Negative
0.040
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
0
5
10
15
20
25
30

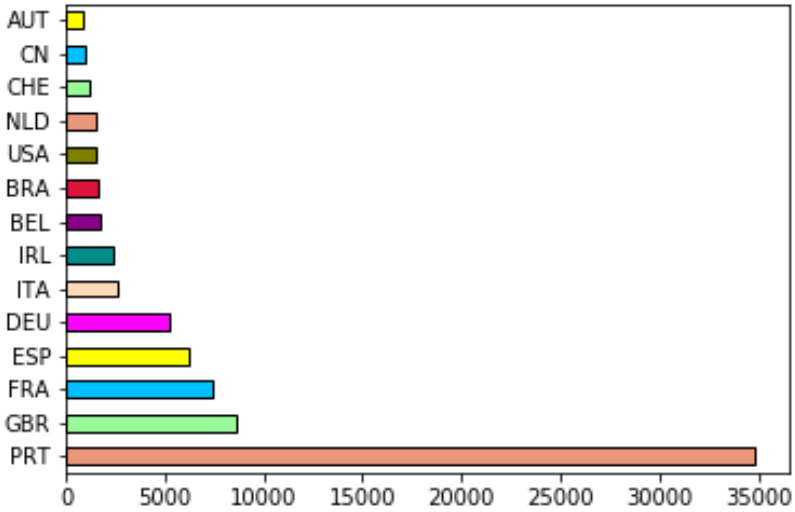

order_month
Positive
Negative
0.10
0.08
0.06
0.04
0.02
0.00
2
4
6
8
10
12

נספח ב': גרפים של Acquisition Channel ו-Order Type

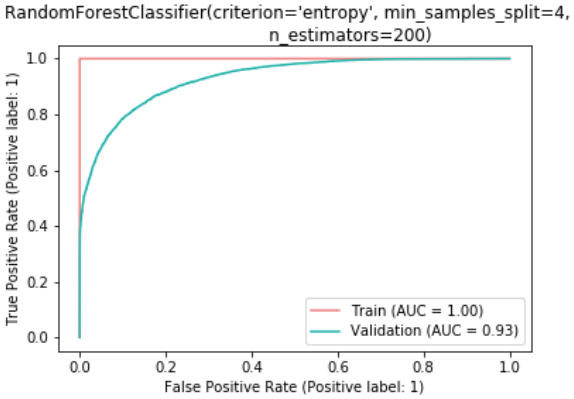




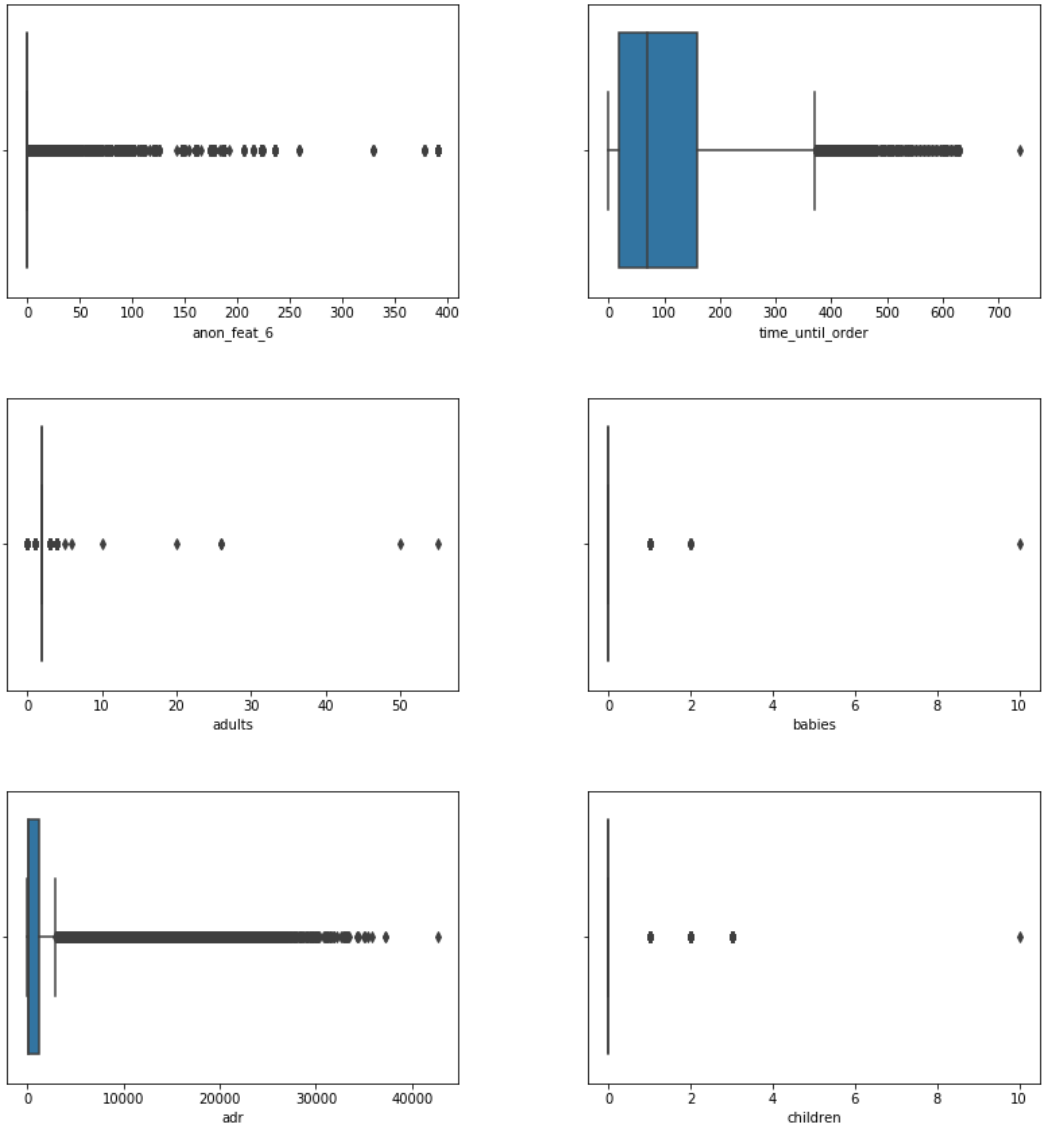
נספח ג': גרף המדינות המופיעות ביותר מ-1% מהדאטה



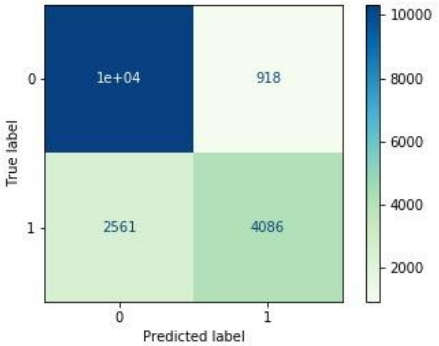
נספח ד': גרף AUC של מודל Random Forest על סט ה-Validation



נספח ה': גרפי Boxplot של פיצ'רים בעלי Outliers



נספח ו': confusion matrix למודל logistic regression על סט ה-validation



נספח ז': Grid Search

In [57]:
```
from sklearn.model_selection import GridSearchCV
parametersOptions = {'activation' : ["logistic", "relu"], 'hidden_layer_sizes' : [(100,), (50, 50), (20, 20, 10, 10, 10)],
                     'batch_size' : [10, 50],
                     'learning_rate_init' : [0.1, 0.01],
                     'max_iter' : [1500]}
GS = GridSearchCV(MLPClassifier(), parametersOptions, cv=3, scoring='roc_auc')
GS.fit(X_train, y_train['cancelation'])
print (GS.best_params_)
```

```
{'activation': 'logistic', 'batch_size': 10, 'hidden_layer_sizes': (50, 50), 'learning_rate_init': 0.01, 'max_iter': 1500}
```

In [58]:
```
CV_rfc = GridSearchCV(RandomForestClassifier(), param_grid = {'n_estimators':[100,150,200],'min_samples_leaf': [1,2,3,4],
                                                              'min_samples_split': [2,3,4],'criterion':['gini','entropy']})
CV_rfc.fit(X_train, y_train['cancelation'])
print(CV_rfc.best_params_)
```

```
{'criterion': 'gini', 'min_samples_leaf': 3, 'min_samples_split': 2, 'n_estimators': 200}
```

In [79]:
```
param_grid = {'n_neighbors': [50,55], 'weights': ['uniform','distance']}
CV_knn = GridSearchCV(knn,param_grid , cv=3, scoring = 'roc_auc')
CV_knn.fit(X_train, y_train['cancelation'])
print(CV_knn.best_params_)
```

```
{'n_neighbors': 50, 'weights': 'distance'}
```

In [81]:
```
grid={"C":[0.001, 0.01, 0.1, 1, 10, 100, 1000], "penalty":["l1","l2"]}# l1 lasso l2 ridge
logreg_cv=GridSearchCV(LR_clf, grid, cv=5, scoring = 'roc_auc')
logreg_cv.fit(X_train,y_train['cancelation'])
print(logreg_cv.best_params_)
```

```
{'C': 1000, 'penalty': 'l2'}
```