פרויקט מסכם

חיזוי ביטולי הזמנת מלונות באמצעות למידת מכונה

מגישות : דניז בישבסקי ואלינור בנגייב
קורס: למידת מכונה
שם המרצה: דור בנק
שם המתרגל: שחף גורן

תקציר מנהלים:

מטרת הפרויקט הינה לחזות ביטולי הזמנות לבתי מלון. על מנת לענות על שאלה זו, המשמעותית לתחום המלונאות, ביצענו ניתוח על נתוני עבר שקיבלנו, שבן היתר כלל: עיבוד של הנתונים ובדיקה של אילו מאפיינים של ההזמנה רלוונטיים ומשמעותיים לחיזוי הביטול, במקביל לויזואליזציה של הנתונים ומתן הסברים על הנעשה. לאחר הרצת מודלים והערכתם, בחרנו במודל Random Forest אשר עונה על השאלה באחוז דיוק גבוה של 0.907. באמצעות מודל זה, ביצענו חיזוי על ביטולן של נתוני הזמנות חדשות שקיבלנו והצלחנו לספק תשובה מהימנה על האם ההזמנה תבוטל או לא.

אקספלורציה:

ראשית, ביצענו אקספלורציה בסיסית על הדאטה.

- בדקנו אילו עמודות הוא מכיל וראינו שהוא מכיל עמודות מספריות ועמודות קטגוריאליות שבהמשך כל סוג נדרש לטיפול שונה (נספח 1.1).
- בדקנו את מצב הערכים החסרים בדאטה וראינו כי יש מספר לא קטן של עמודות עם ערכים חסרים. נטפל בנושא בשלב העיבוד המקדים (נספח 1.2).
- בסקירת העמודות ראינו כי לעמודה הראשונה אין שם. הדפסנו את ערכיה וראינו שכולם חד חד ערכיים ולכן הסקנו כי מדובר בעמודת מספר הזמנה ולכן קראנו לעמודה זו 'order_id'. שינינו את שמה בשאר קבצי הנתונים (סט הטסט והלייבל).
- לאחר מכן עברנו לחקור את העמודות המספריות והקטגוריאליות בנפרד -
  - Numerical features exploration -
    - הדפסנו גרפים של כל עמודות אלו כדי לראות את ההתפלגות שלהן, מצאנו כי הן מתפלגות על סקיילים (scales) שונים וראינו שהרבה מהן מוטות (נספח 1.3).
    - יצרנו גרף המראה את כמות הביטולים לעומת כמות הצ'ק-אין (לא בוטל) של כל ההזמנות. ראינו כי 37% מההזמנות בוטלו, זה מצביע על כך שהדאטה מוטה במקצת ולא מאוזן (נספח 1.4).
    - חקירת נתונים חריגים - ע"י השוואה בין ערכי הממוצע, המינימום והמקסימום של כל פיצ'ר, מצאנו אילו מהם מכילים ערכים חריגים.
    - בדיקת קורלציה - יצרנו heat map של מתאמי פירסון בין הפיצ'רים המספריים ומצאנו מתאם גבוה בין כמה מהם (נספח 1.5).
  - Categorical features exploration -
    - יצירת גרפים המתארים את התפלגות העמודות הקטגוריאליות והצגת 10 המשתנים הפופולריים ביותר בכל עמודה. ראינו כי העמודות order type ו-acquisition channel הן בעלות קטגוריות דומות לכן אולי הן מתארות דברים דומים (נספח 1.6).
    - יצרנו גרף אינטראקטיבי המראה על כדור הארץ את אחוז הביטולים עבור כל מדינה ביחס לכמות ההזמנות שלה, מה שמאפשר לנו לראות באילו מדינות אחוז זה הכי גבוה (נספח 1.7).

עיבוד מקדים:

טיפול בערכים חסרים:

הדאטה הכיל ערכים חסרים הן במשתנים הקטגוריאליים והן במשתנים המספריים (נספח 2.1). טיפול בערכים חסרים חשוב על מנת שנוכל להריץ פעולות ומודלים בצורה תקינה ולכן השתמשנו בטכניקות שונות כדי לטפל בבעיה זו:

1. תחילה השמטנו פיצ'רים אשר כמות הערכים החסרים בה הייתה מעל ל90%. לא ניתן למלא אותם בערכים אחרים כיוון שזו תהיה התערבות גסה מידי מצידנו שעשויה להטות את המודל.
2. הסרנו פיצ'רים בעלי משמעות זהה עם פיצ'רים ללא ערכים חסרים. לדוגמה, לפיצ'ר order_month היו כ-3.8 אחוז ערכים חסרים, אך לעומת זאת הפיצ'ר order_week לא הכיל ערכים חסרים כלל. כיוון שיכולנו לחשב מה הם החודשים שבהם נעשו ההזמנות בעזרת מספר השבועות העדפנו להסיר את order_month מאשר למלא את ערכיו החסרים.
3. מילאנו את הערכים החסרים וה-undefined של שאר הפיצ'רים ע"י חציון, והערך הנפוץ ביותר. את ההחלטה באיזו מתודה להשתמש קבענו לפי ההתפלגות הפיצ'רים שראינו בשלב האקספלורציה. בנוסף, בנקודה זו פיצלנו את הדאטה לtrain ו- validation כיוון שלולא היינו עושים זאת, היינו מלמדים את האלגוריתם לחשב את החציון והשכיח גם לפי הוולידציה שלנו וזה אסור בהחלט.
   - עבור פיצ'רים מספריים בעלי התפלגות רציפה ומוטה השתמשנו בחציון, מכיוון שהם מכילים outliers רבים ושימוש בממוצע עשוי להטות את ההתפלגות.
   - עבור פיצ'רים קטגוריאליים, או מספריים בעלי התפלגות בדידה השתמשנו בערך הנפוץ ביותר.

כמו כן, מילאנו את הערכים החסרים בvalidation וב-test לפי ערכים שחושבו על פי ה-train.

טיפול ב Outliers:

הדאטה שקיבלנו מכיל משתנים מוטים בעלי זנבות ארוכים של Outliers (נספח 2.2). לאחר ניסיונות טיפול בהם באמצעות מתודת z-score, ראינו שגם עבור 5 סטיות תקן אנו מורידים כמות דאטה גדולה מדי, דבר שפגע בביצועי המודל ולכן החלטנו להשאיר את ה-outliers. כדי לטפל בבעיית הזנבות הארוכים השתמשנו בטרנספורמציית ה-standardization עליה נרחיב בסעיף ה- Data scaling.

קידוד משתנים קטגוריאליים:

בחרנו להשתמש בקידוד מסוג "get_dummies" ההופך משתנים קטגוריאליים למשתנים מספריים מלאכותיים בשם dummy variables. סוג הקידוד המקביל לו הוא ה"label encoding". הסיבה שלא בחרנו בשיטה זו היא מכיוון שהקידוד של המשתנים מתבצע לפי סדר אלפביתי מ0 עד מספר הקטגוריות, ולכן יכול להיווצר מצב שבו המודל ישער ששני ערכים קרובים הם יותר דומים מאשר שני ערכים רחוקים יותר, כאשר בפועל אין יחס סדר בין הקטגוריות, דבר שלא מתאים למקרה שלנו. לעומת זאת, השימוש ב dummy variables שומר על אי תלות בין המשתנים וזה התאים יותר לסוג המשתנים שהדאטה שלנו הכיל. החיסרון העיקרי בdummy variables הוא הגדלת המימדיות אך בבעיה זו טיפלנו בהמשך.

Data scaling:

בשלב האקספלורציה ראינו כי הדאטה אינו מנורמל (נספח 2.3), וכי הפיצ'רים בעלי שונות גדולה, דבר שיכול לפגוע בביצועי המודל. מעבר לכך, שלב הscaling הוא שלב חשוב כיוון שאלגוריתמים מסוימים לא יודעים להתמודד עם דאטה בעל הבדלים גדולים בscale. ישנם שתי מתודות נפוצות ליצירת אחידות בסקיילים של הדאטה ולהקטנת השונות הן - Min Max Scaling (נורמליזציה) ו- Standardization. בחרנו להשתמש ב Standardization כיוון שהוא יותר רובסטי עבור משתנים בעלי Outliers מרובים. לעומתו Min Max Scaling מושפע מאוד מערכים חריגים, והוא בעצם מבטל אותם, וכן יש להתייחס לערכים אלו במודל כיוון שלא הורדנו אותם.

חשוב לציין שבמהלך שלב העיבוד המקדים ה-scaling נעשה אך ורק לפי הtrain set והתוצאות שלו יושמו על גבי הvalidation set והtest set. כמו כן, יישמנו את ה-scaling רק על גבי המשתנים המספריים שאינם משתני dummy.
כדי לוודא שה-scaling הצליח בדקנו שהממוצע של הפיצ'רים עליהם יישמנו את המתודה היו שווה ל-0 ושסטיית התקן שווה ל- 1.

טיפול בבעיית המימדיות:

ריבוי מימדים מתרחש כאשר מספר הפיצ'רים שאנו משתמשים בו גדול וזה עשוי ליצור בעיה מכיוון שחלק מאלגוריתמי החישוב של המודלים נעזרים בחישוב מרחק אוקלידי. ככל שיש יותר מימדים (פיצ'רים), המרחק בין דגימות עולה ולכן הן יהיו רחוקות אחת מהשניה. ברגע שנרצה לבצע פרידקציה על דגימה חדשה, יכול להיות מצב שהיא תהיה מאוד רחוקה מדגימות האימון ולכן הפרדיקציה תהיה פחות אמינה, ולכן בסופו של דבר נקבל מודל שהוא overfitted. בנוסף, ככל שגודל רמת המרחבים שיש לחשב עולה, החישוב הופך להיות לא יעיל הן מבחינת כוח וזמן החישוב.
כדי להקטין את המימדיות השתמשנו בFeature selection וPCA.

Feature selection:

1. בעזרת בדיקת קורלציה בדקנו אילו פיצ'רים נמצאים בקורלציה הכי נמוכה ביחס ללייבל ואותם הסרנו (נספח 2.4).
2. שמנו לב שהמשתנים של acquisition_channel היו דומים מאוד לאלו של order_type (נספח 2.5). בעזרת במבחן חי בריבוע קבענו ברמת ביטחון של 95% שהפיצ'רים תלויים ולכן הסרנו את הראשון, משום של-order_type יש יותר קטגוריות (דבר שתורם להסבר שונות הדאטה).
3. בנינו פיצ'רים חדשים, המאפשרים לנו להסיר משתנים אחרים, ע"י מניפולציות ופעולות מתמטית על הפיצ'רים קיימים:
   - הסרנו את הפיצ'ר country ובמקומו יצרנו פיצ'ר בשם is_in_top_ten_countries ובו הצבנו 1 במידה והמדינה נמצאה בעשרת המדינות הפופולריות ביותר ו- 0 אם לא.
   - הסרנו את order_week ובמקומו יצרנו פיצ'ר בשם season שחילק את השבועות לפי עונות.
   - הסרנו את order_day_of_month ויצרנו פיצ'ר בשם time_in_month שסיווג את זמן ההזמנה לתחילת החודש, אמצע החודש וסוף החודש.
   - מיזגנו את העמודות מבוגרים, ילדים ותינוקות לפיצ'ר בשם num_of_guests והסרנו עמודות אלה.
   - יצרנו פיצ'ר בשם cancellation_rate שמדד את אחוז הביטולים של הלקוח והסרנו את הפיצ'רים prev_canceled ו- prev_not_canceled.

PCA:

- לאחר שיצרנו את משתני ה- dummy החזקנו ב 42 פיצ'רים ולכן נעזרנו ב PCA. בעזרת גרף (נספח 2.6) יכולנו לבדוק עד כמה הפיצ'רים הסבירו את השונות של הדאטה שלנו. ראינו שמתוך 42, 28 פיצ'רים יכלו להסביר 99% מהשונות ולכן צימצמנו בעזרת הPCA את כמות הפיצ'רים ל 28. חשוב לציין שצמצום זה לא יחליש באופן מהותי את יכולת החישוביות של המודל אבל זה כן יטיב איתנו
- בעת בחירת ההיפר-פרמטרים האופטימיליים למודלים בשלב הבא.

החלת העיבוד המקדים על ה-Test:

בסוף שלב העיבוד המקדים הרצנו את כל הפעולות שעשינו על גבי ה-test set לפי הערכים שחישבנו **על ה-train set,** לדוגמה הממוצע, סטיות התקן וכדומה.

הרצת המודלים

בחנו להריץ 4 מודלים שונים ולבדוק מי יניב את תוצאת ה-AUC הגבוהה ביותר.

- עבור כל מודל נציג את ההיפר-פרמטרים שנבחרו ע"י שימוש במתודת Grid Search CV, המאפשרת לבחור את הפרמטרים האופטימליים עבור מודל ספציפי מתוך סט פרמטרים שהגדרנו לה לבדוק.
- לאחר מכן נציג את הממוצע של ה- AUC עבור 10 folds שחושבו בעזרת מתודת הk-fold cv שתציג עד כמה המודל יציב וכמו כן האם הוא overfitted או underfitted.

מודלים ראשוניים:

Logistic Regression:

להלן ההיפר פרמטרים האופטימליים שקיבלנו:

```
Fitting 3 folds for each of 144 candidates, totalling 432 fits

{'C': 0.01, 'max_iter': 50, 'penalty': 'l2', 'solver': 'lbfgs', 'tol': 0.0001}
```

הAUC שהתקבל בעזרת פרמטרים אלו על סט הvalidation הינו 0.8507 ובגרף ה k-fold cv ה-AUC הממוצע שהתקבל הינו 0.8475 (נספח 3.1).

KNN:

להלן ההיפר פרמטרים האופטימליים שקיבלנו:

```
Fitting 3 folds for each of 16 candidates, totalling 48 fits

{'metric': 'manhattan', 'n_neighbors': 20, 'weights': 'distance'}
```

הAUC שהתקבל בעזרת פרמטרים אלו על סט הvalidation הינו 0.8978 ובגרף ה k-fold cv ה-AUC הממוצע שהתקבל הינו 0.8978 (נספח 3.2).

מודלים מתקדמים:

Decision Tree:

להלן ההיפר פרמטרים האופטימליים שקיבלנו:

```
Fitting 3 folds for each of 18 candidates, totalling 54 fits

{'criterion': 'gini',
 'max_depth': None,
 'min_samples_leaf': 4,
 'min_samples_split': 2}
```

הAUC שהתקבל בעזרת פרמטרים אלו על סט הvalidation הינו 0.7971 ובגרף ה k-fold cv ה-AUC הממוצע שהתקבל הינו 0.8010 (נספח 3.3).

Random Forest:

להלן ההיפר פרמטרים האופטימליים שקיבלנו:

```
Fitting 3 folds for each of 162 candidates, totalling 486 fits

{'bootstrap': True,
 'criterion': 'entropy',
 'max_depth': None,
 'max_features': 'auto',
 'min_samples_leaf': 2,
 'min_samples_split': 2,
 'n_estimators': 250}
```

הAUC שהתקבל בעזרת פרמטרים אלו על סט הvalidation הינו 0.9074 ובגרף ה k-fold cv ה-AUC הממוצע שהתקבל הינו 0.9092 (נספח 3.4).

הערכת המודלים

בשלב זה, לאחר שהרצנו את כל המודלים הערכנו אותם ע"י AUC. הרצנו K-Fold Cross Validation, עם 10 פולדים (K=10) וקיבלנו עבור כל מודל ציון AUC ממוצע. בחרנו להריץ את הk fold על כל הדאטה משום שיותר דאטה משמע אימון יותר טוב וביצועים יותר טובים.
בנוסף, ביצענו פרדיקציה על סט הvalidation וחישבנו ציון AUC עבור כל מודל. בחרנו במודל בעל ציון הAUC הגבוה ביותר שהוא Random Forest.
בחינת overfitting - ראשית, על מנת להגדיל את יכולת הכללה של המודל, בחרנו את ההיפר פרמטרים האופטימליים בשלב הרצת המודלים. שנית, משמעות overfitting היא שהמודל לא מוכלל עבור סטים שונים של נתונים ולכן אם המודל overfitted, עבור כל סט הוא יתנהג אחרת וביצועיו יהיו שונים. במקרה שלנו, ניתן לראות בגרף כי העקומות של הפולדים מאוד קרובות והשונות ביניהם נמוכה. דבר זה מעיד על כך שהמודל שלנו יציב ושהוא מתנהג דומה עבור סטים של נתונים שונים. לכן, ניתן להסיק כי המודל מכליל בצורה טובה ולכן הוא לא overfitted. כמו כן, ביצענו פרדיקציה על סט הvalidation, וחישבנו את ציון הAUC. ראינו שהפער בינו לבין ציון הAUC הממוצע בK-Fold מאוד קטן ולכן ניתן שוב לקבוע שהמודל לא overfitted.

Confusion Matrix on Random Forest Model

נוכל לראות את המטריצה בנספח 3.5. בחרנו לבנות את המטריצה על המודל הנבחר.
משמעות התאים היא -
TP - תא שמאלי עליון - 10792 הזמנות אשר המודל חזה שלא בוטלו, לא בוטלו. **סיווג נכון**
FN - תא שמאלי תחתון - 1042 הזמנות אשר המודל חזה שלא בוטלו, כן בוטלו. **סיווג לא נכון**
FP - תא ימני עליון - 416 הזמנות אשר המודל חזה שבוטלו, לא בוטלו באמת. **סיווג לא נכון**
TN - תא ימני תחתון - 5659 הזמנות אשר המודל חזה שבוטלו, באמת בוטלו. **סיווג נכון**
חישבנו ציוני presicion ו recall, וקיבלנו -
precision score 0.96 - כלומר, הפרדיקציה שלנו עבור הרוב המוחלט של התצפיות נכונה.
recall score 0.91 - על פי ציון זה, המודל טעון שיפור, משום שעבור לא מעט תצפיות אשר במציאות מסווגות כtrue, המודל מפספס ומסווג אותן כfalse.

סיכום

מטרת הפרויקט הינה לחזות ביטולי הזמנות לבתי מלון. קיבלנו קובץ דאטה המכיל מאפיינים עבור כל הזמנה, חלקם בעלי משמעות ידועה וחלקם לא. ביצענו אקספלורציה על הדאטה, תחילה על כולו ואז עברנו לחקור את העמודות המספריות והעמודות הקטגוריאליות בנפרד, הסקנו בשלב זה משמעויות שונות שאותן יישמנו בשלב העיבוד המקדים על הדאטה. בשלב זה, טיפלנו בערכים חסרים, קידדנו משתנים בעמודות הקטגוריאליות, יצרנו פיצ'רים חדשים מתוך הפיצ'רים הקיימים, וביצענו scaling על העמודות המספריות. בנוסף בשלב זה, התמודדנו עם בעיית המימדיות, הורדנו פיצ'רים אשר לא תורמים כמעט דבר עבור החיזוי, וכן השתמשנו במודל PCA אשר צימצם לנו את כמות הפיצ'רים. כל זאת על מנת לאפשר למודלים ללמוד את הנתונים בצורה מיטבית על מנת שהפרידקציה תהיה מדויקת ככל הניתן. בהמשך, יצרנו מודלים אשר עבור כל אחד ביצענו הערכה של טיב המודל באמצעות K-Fold-Cross Validation, בחרנו את המודל Random Forest אשר ביצועיו הכי גבוהים, בעל ציון auc = 0.907. לבסוף, ביצענו פרדיקציה עבור סט הנתונים החדש (סט הטסט), שאותה ייצאנו לקובץ חדש, שמציג את ההסתברות לביטול עבור כל הזמנה.

נספחים:

1. נספחי אקספלורציה:

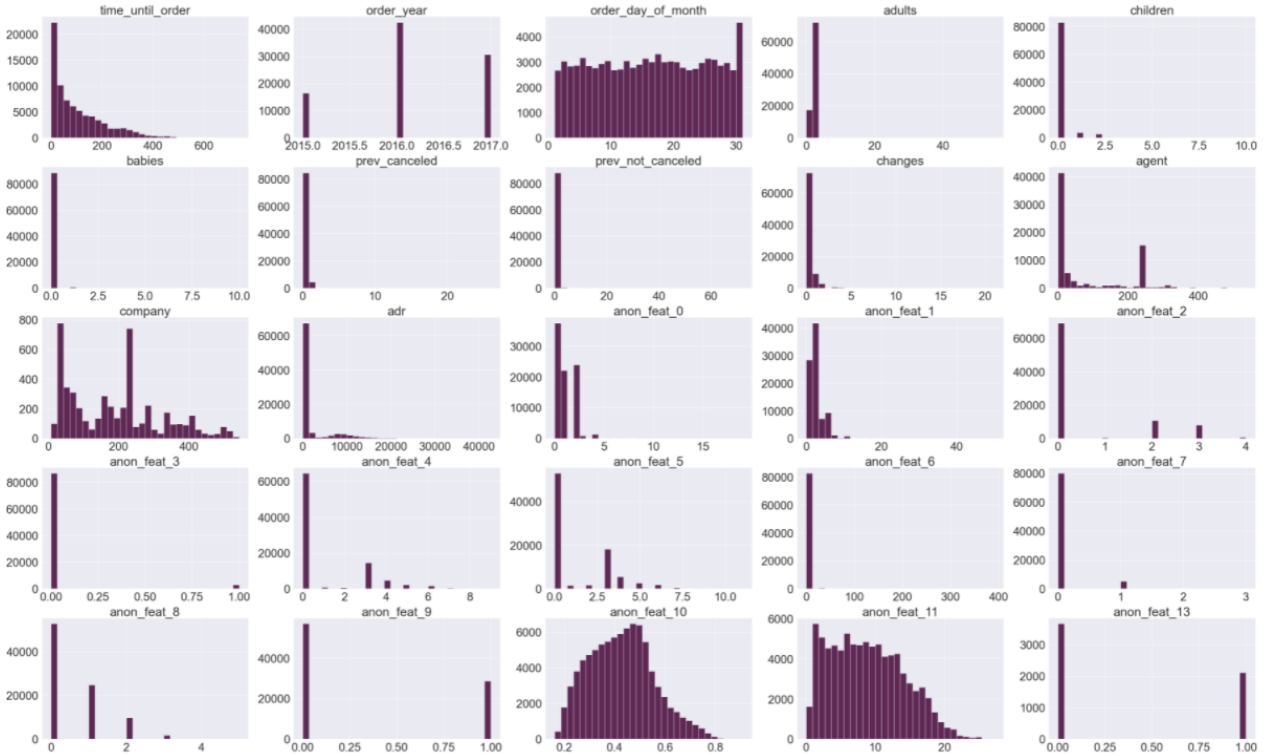
1.1 - סוגי העמודות בדאטה

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 89542 entries, 0 to 89541
Data columns (total 34 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Unnamed: 0            89542 non-null  int64
 1   time_until_order      76861 non-null  float64
 2   order_year            89542 non-null  int64
 3   order_month           86108 non-null  object
 4   order_week            89542 non-null  object
 5   order_day_of_month    89542 non-null  int64
 6   adults                89542 non-null  int64
 7   children              89538 non-null  float64
 8   babies                89542 non-null  int64
 9   country               85201 non-null  object
 10  order_type            89542 non-null  object
 11  acquisition_channel   89542 non-null  object
 12  prev_canceled         89542 non-null  int64
 13  prev_not_canceled     89542 non-null  int64
 14  changes               86065 non-null  float64
 15  deposit_type          80536 non-null  object
 16  agent                 77346 non-null  float64
 17  company               5062 non-null   float64
 18  customer_type         79647 non-null  object
 19  adr                   86559 non-null  float64
 20  anon_feat_0           86161 non-null  float64
 21  anon_feat_1           89542 non-null  int64
 22  anon_feat_2           89542 non-null  int64
 23  anon_feat_3           89542 non-null  int64
 24  anon_feat_4           89542 non-null  int64
 25  anon_feat_5           85510 non-null  float64
 26  anon_feat_6           85309 non-null  float64
 27  anon_feat_7           85294 non-null  float64
 28  anon_feat_8           89542 non-null  int64
 29  anon_feat_9           85811 non-null  float64
 30  anon_feat_10          86810 non-null  float64
 31  anon_feat_11          84585 non-null  float64
 32  anon_feat_12          89542 non-null  bool
 33  anon_feat_13          5776 non-null   float64
dtypes: bool(1), float64(14), int64(12), object(7)
memory usage: 22.6+ MB
```

1.2 - ערכים חסרים בדאטה

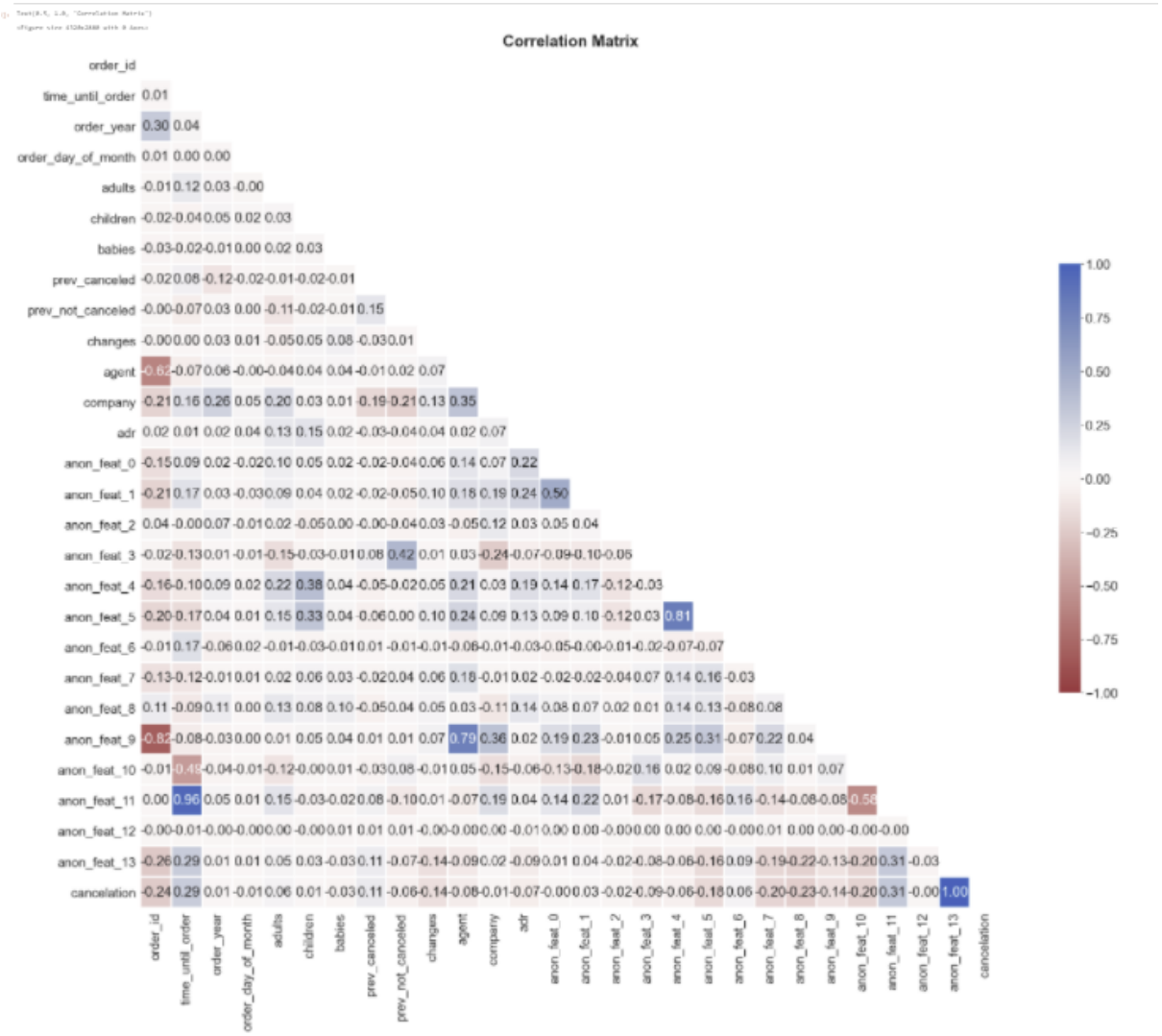
```
[7]: Unnamed: 0              0.000000
     time_until_order       14.162069
     order_year              0.000000
     order_month             3.835072
     order_week              0.000000
     order_day_of_month      0.000000
     adults                  0.000000
     children                0.004467
     babies                  0.000000
     country                 4.848004
     order_type              0.000000
     acquisition_channel     0.000000
     prev_canceled           0.000000
     prev_not_canceled       0.000000
     changes                 3.883094
     deposit_type           10.057850
     agent                  13.620424
     company                94.346787
     customer_type          11.050680
     adr                     3.331398
     anon_feat_0             3.775882
     anon_feat_1             0.000000
     anon_feat_2             0.000000
     anon_feat_3             0.000000
     anon_feat_4             0.000000
     anon_feat_5             4.502915
     anon_feat_6             4.727390
     anon_feat_7             4.744142
     anon_feat_8             0.000000
     anon_feat_9             4.166760
     anon_feat_10            3.051082
     anon_feat_11            5.535950
     anon_feat_12            0.000000
     anon_feat_13           93.549396
     dtype: float64
```

1.3 - אקספלורציה של הפיצ'רים המספריים



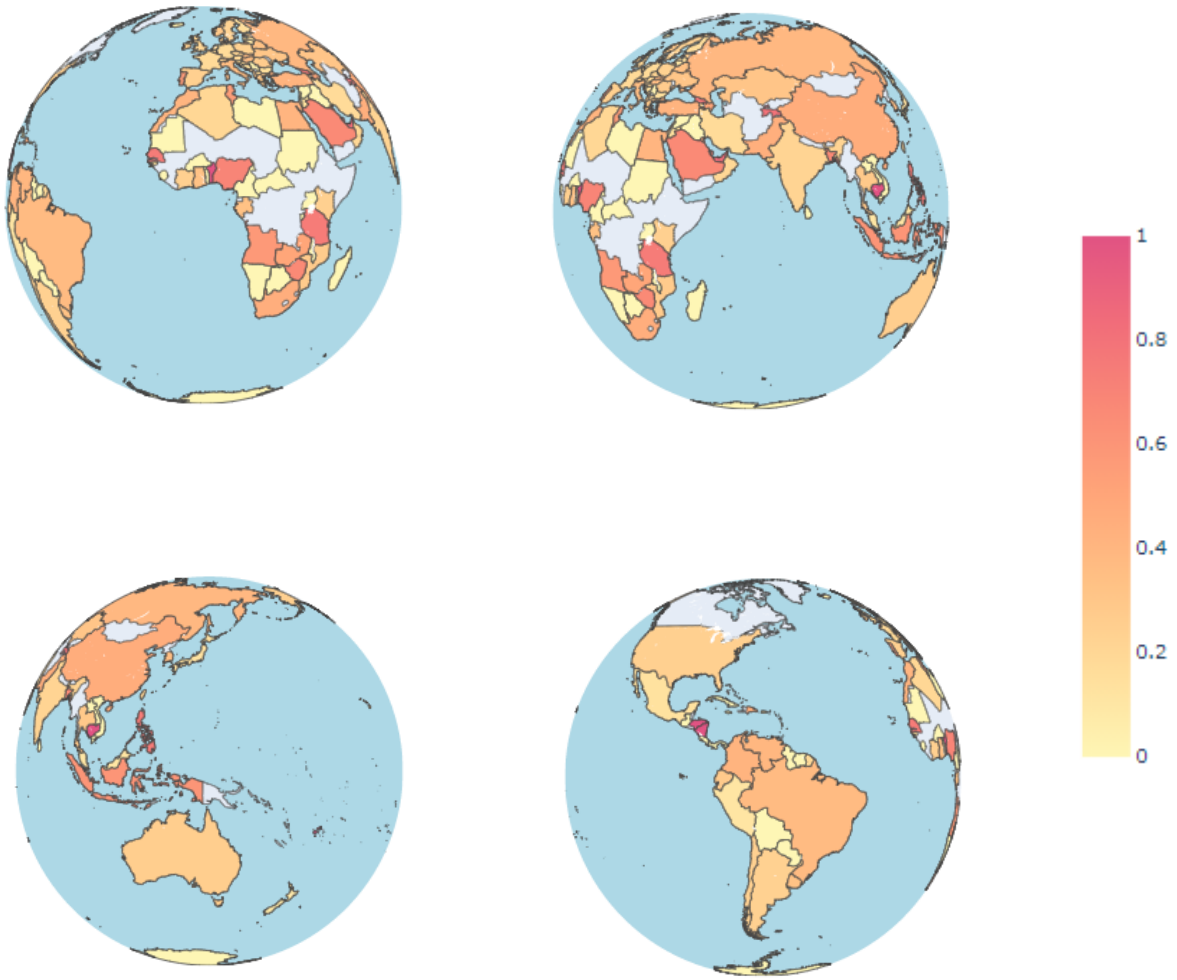
1.4 - גרף של כמות הביטולים לעומת כמות הצ'ק אין

1.5 - גרף heat map עבור מתאמי פירסון בין הפיצ'רים



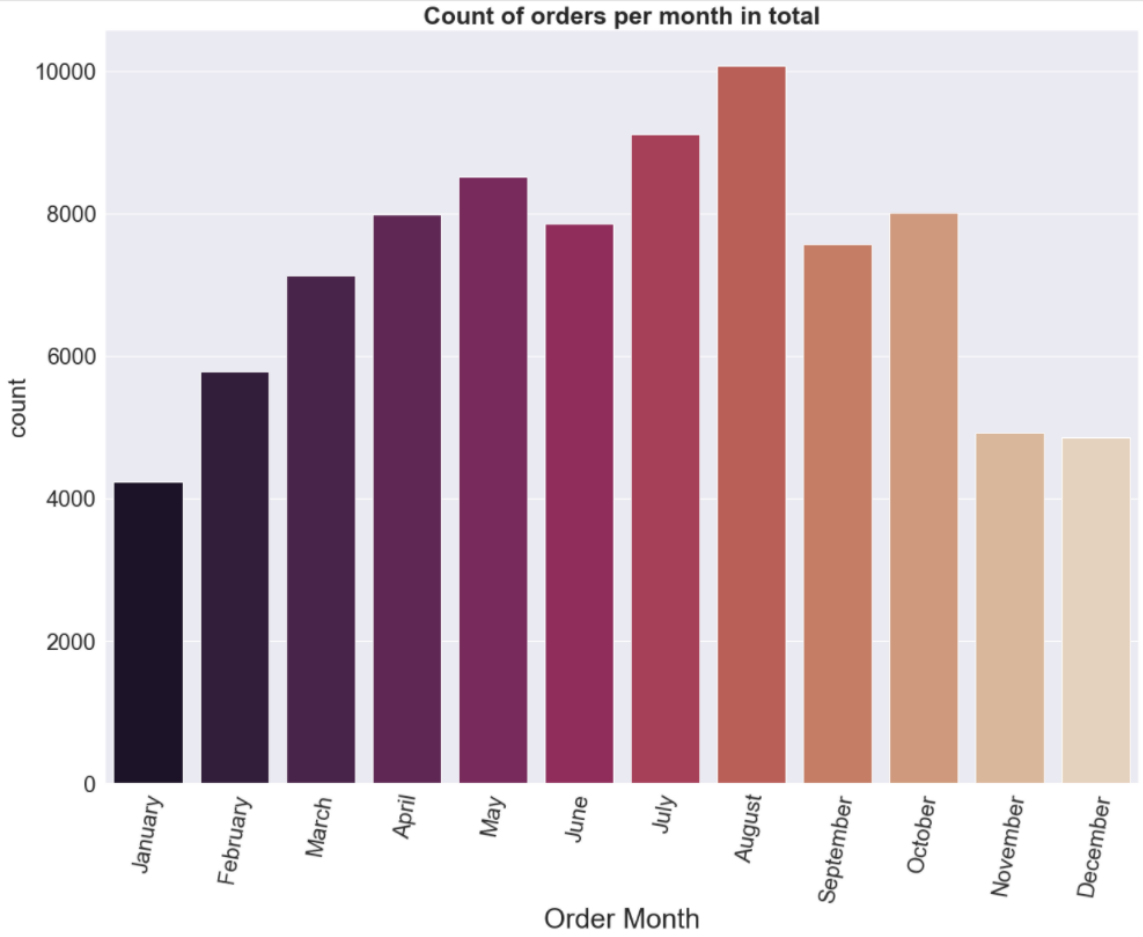
1.6 - אקספלורציה של הפיצ'רים הקטגוריאליים

1.7 - גרף אינטראקטיבי של אחוז הביטולים עבור כל מדינה ביחס לכמות הזמנותיה

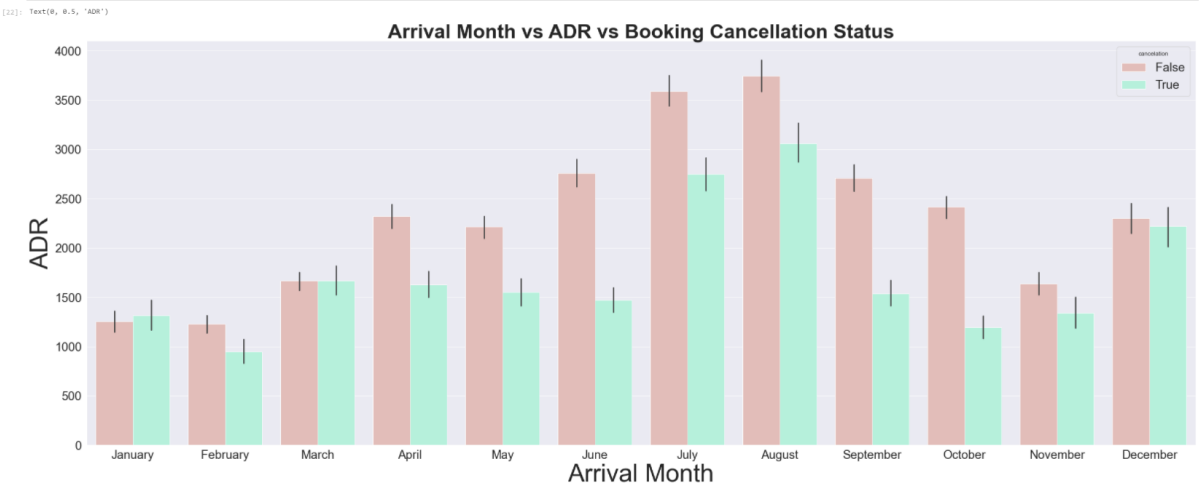


גרפים נוספים:

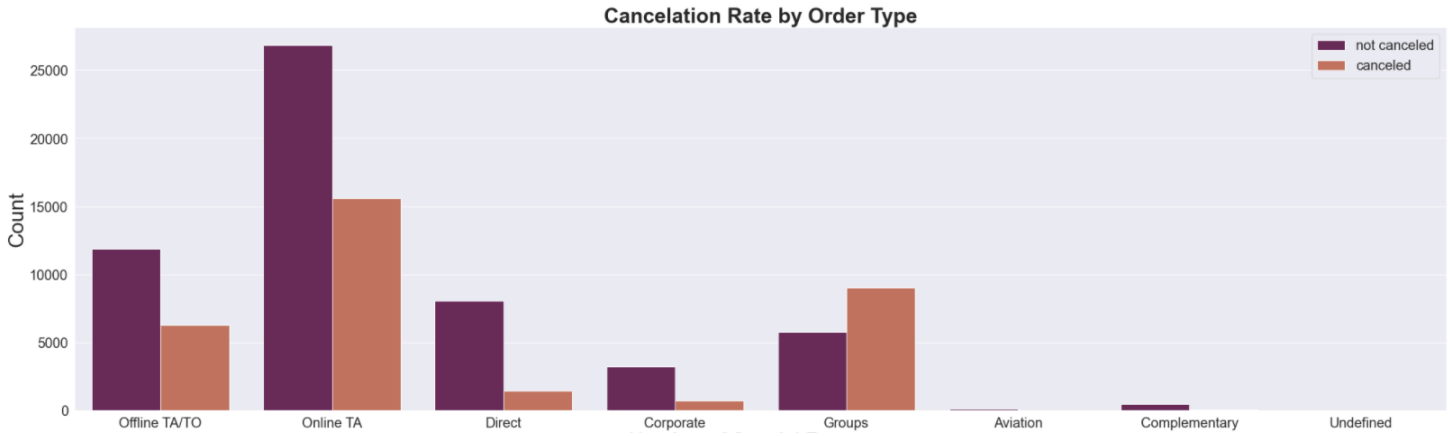
1.6 - גרף מספר הזמנות פר חודש



1.7 - גרף Month vs ADR vs Cancelation Status



1.8 גרף ביטולים לפי סוג הזמנה

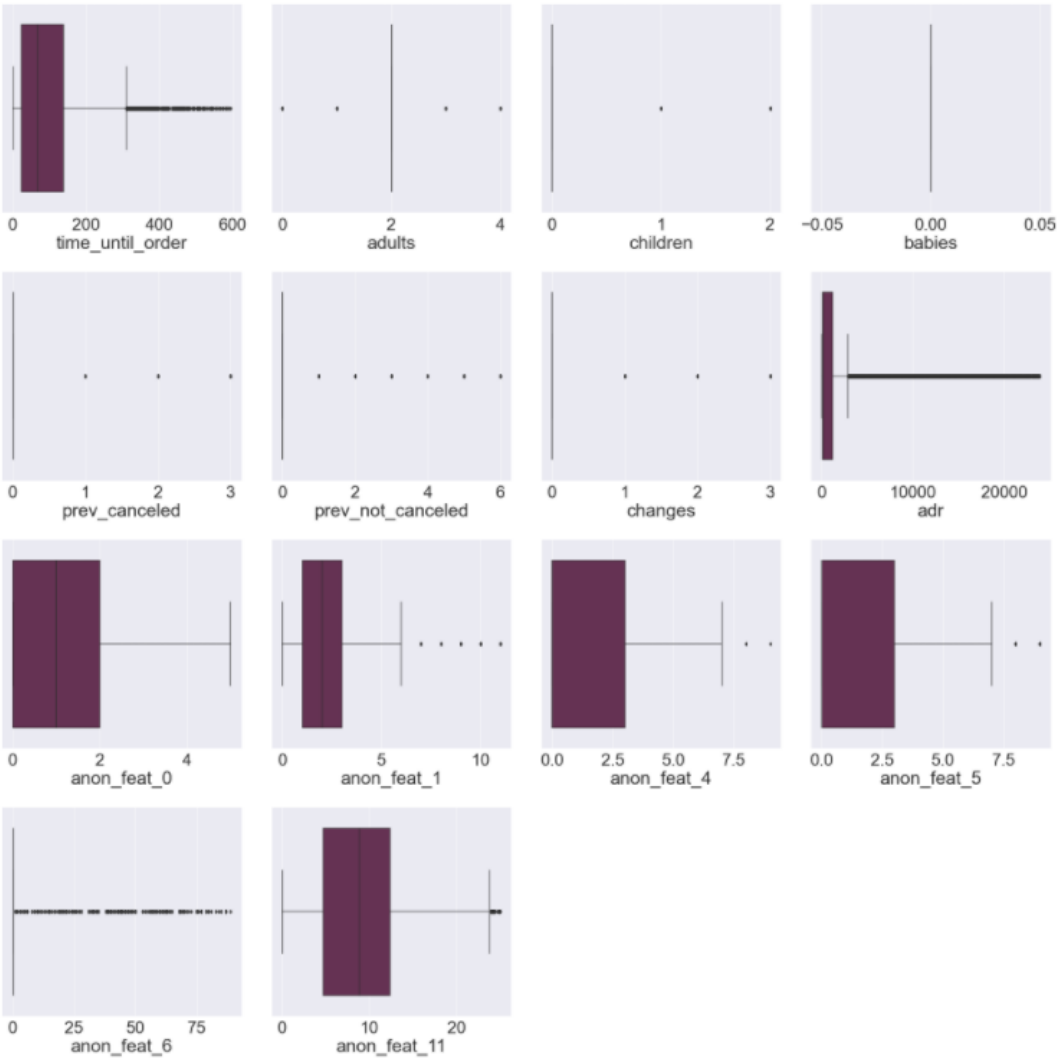


2. נספחי עיבוד מקדים

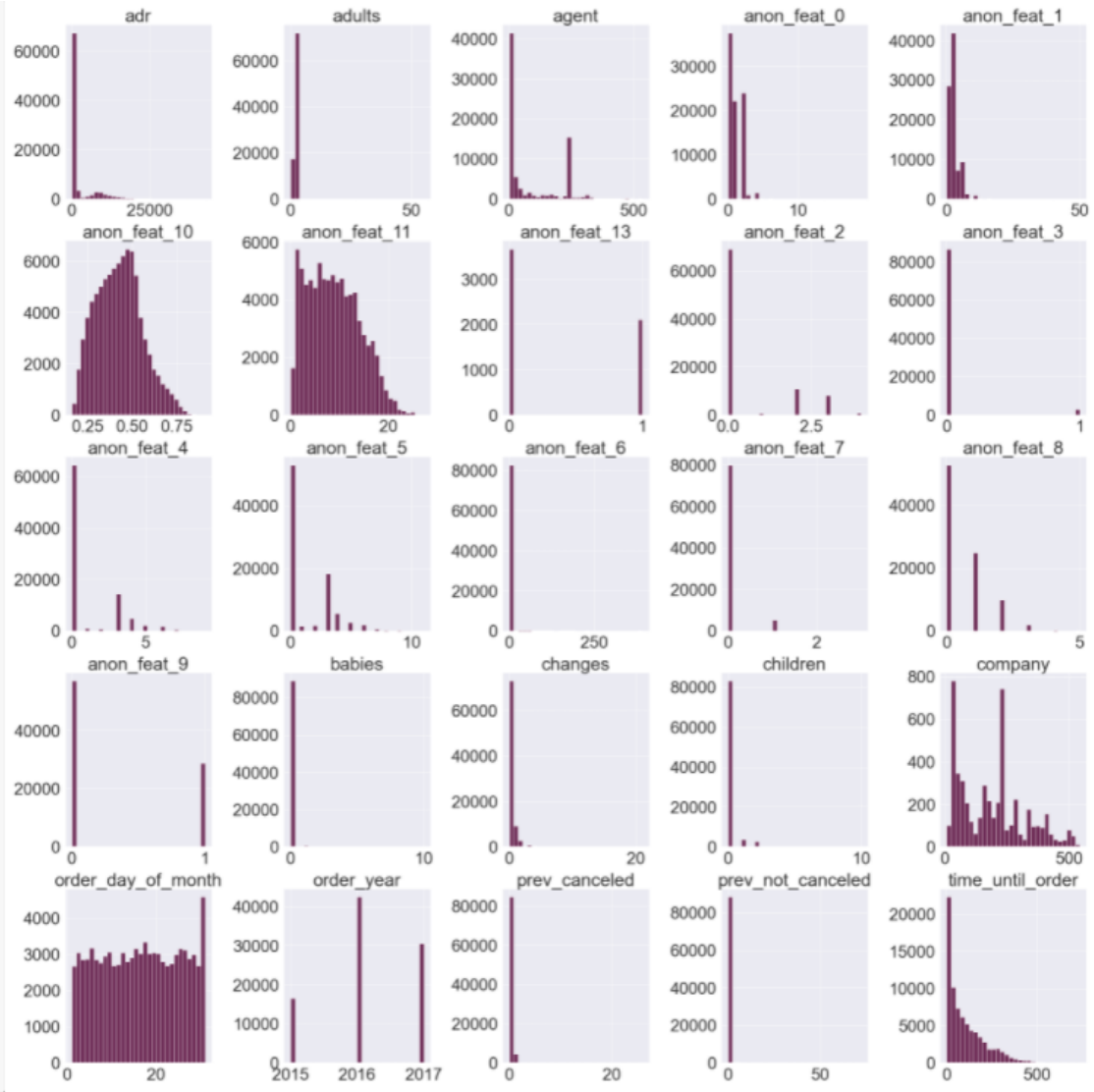
2.1 ערכים חסרים בדאטה:

```
order_id                0.000000
time_until_order       14.162069
order_year              0.000000
order_month             3.835072
order_week              0.000000
order_day_of_month      0.000000
adults                  0.000000
children                0.004467
babies                  0.000000
country                 4.848004
order_type              0.000000
acquisition_channel     0.000000
prev_canceled           0.000000
prev_not_canceled       0.000000
changes                 3.883094
deposit_type           10.057850
agent                  13.620424
company                94.346787
customer_type          11.050680
adr                     3.331398
anon_feat_0             3.775882
anon_feat_1             0.000000
anon_feat_2             0.000000
anon_feat_3             0.000000
anon_feat_4             0.000000
anon_feat_5             4.502915
anon_feat_6             4.727390
anon_feat_7             4.744142
anon_feat_8             0.000000
anon_feat_9             4.166760
anon_feat_10            3.051082
anon_feat_11            5.535950
anon_feat_12            0.000000
anon_feat_13           93.549396
cancelation             0.000000
dtype: float64
```

2.2 פיצ'רים בעלי Outliers רבים



2.3 לפיצ'רים יש scaling שונים

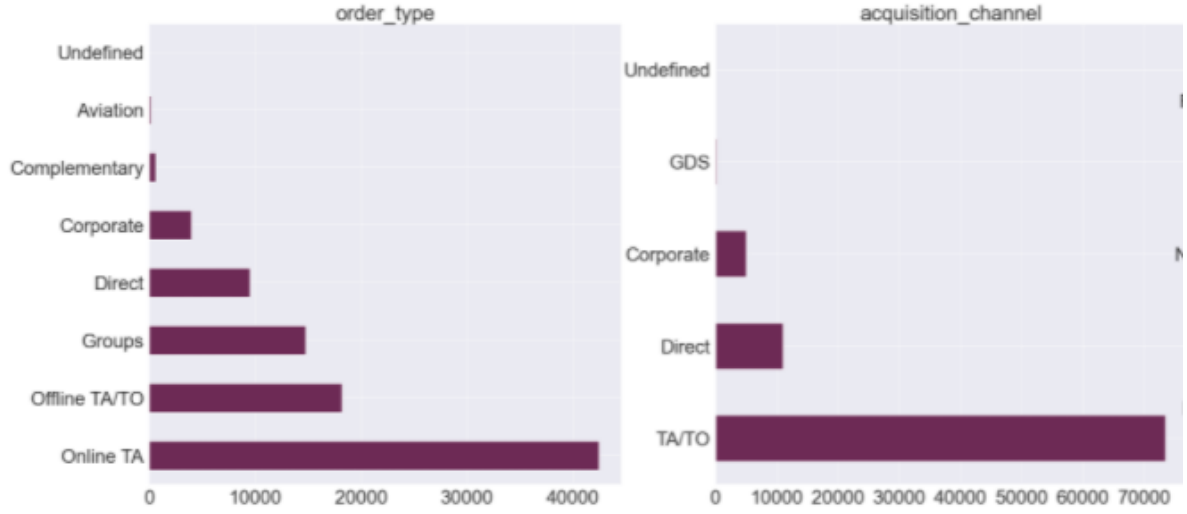


2.4 קורלציה של הפיצ'רים ביחס לליבל (לפני יישום get_dummies):

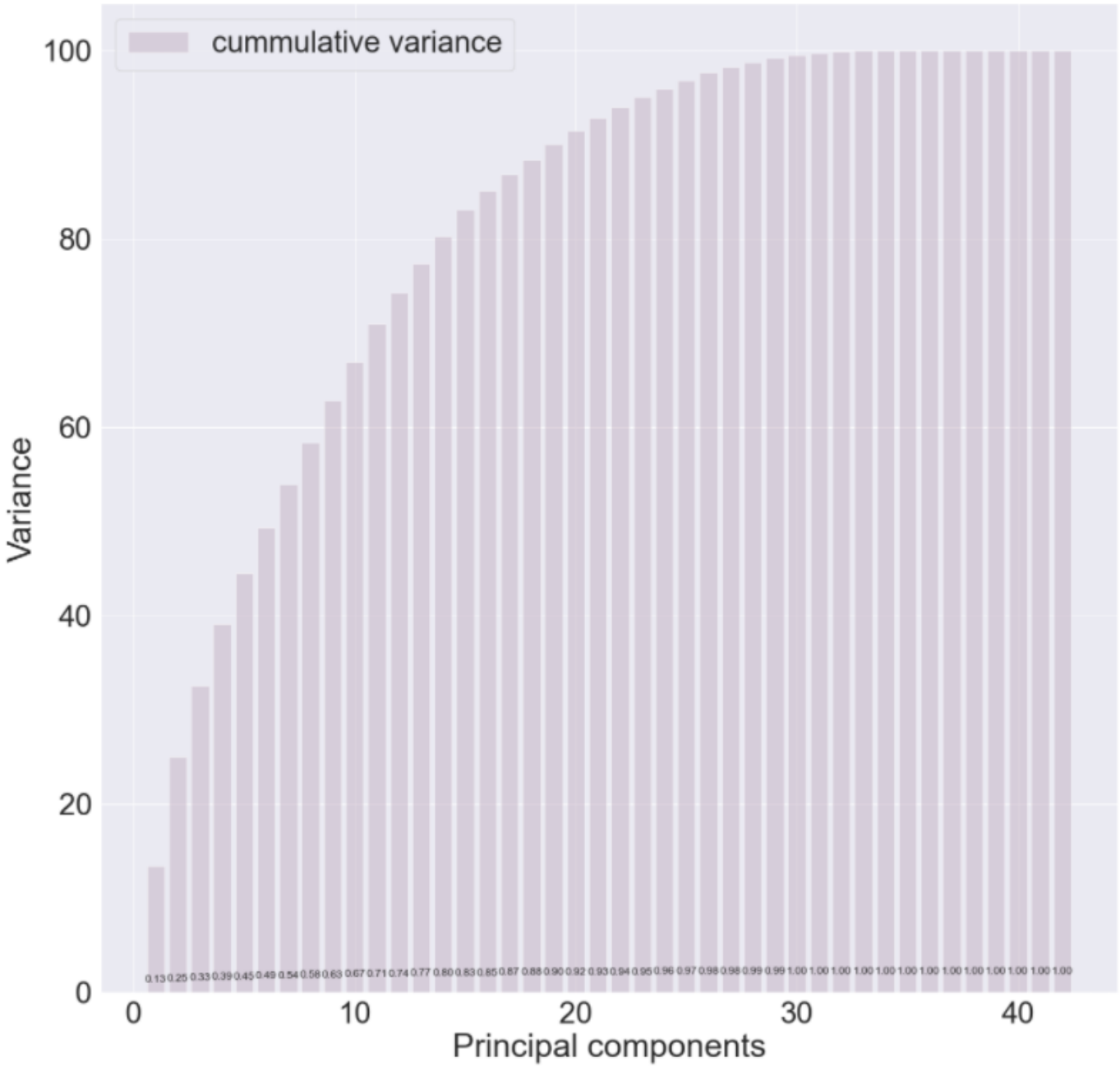
```
numerical_features.corrwith(y_train).abs().sort_values(ascending=False)
```

```
anon_feat_11         0.31
time_until_order     0.29
anon_feat_8          0.23
cancellation_rate    0.23
anon_feat_10         0.20
anon_feat_7          0.20
anon_feat_5          0.18
changes              0.14
anon_feat_9          0.14
prev_canceled        0.11
anon_feat_3          0.09
agent                0.08
adr                  0.07
anon_feat_4          0.06
prev_not_canceled    0.06
anon_feat_6          0.06
adults               0.06
num_of_guests        0.04
babies               0.03
anon_feat_1          0.03
anon_feat_2          0.02
order_week           0.01
order_day_of_month   0.01
anon_feat_12         0.01
children             0.01
anon_feat_0          0.00
dtype: float64
```

2.5 **השוואת המשתנים acquisition_channel ו- order_type:**

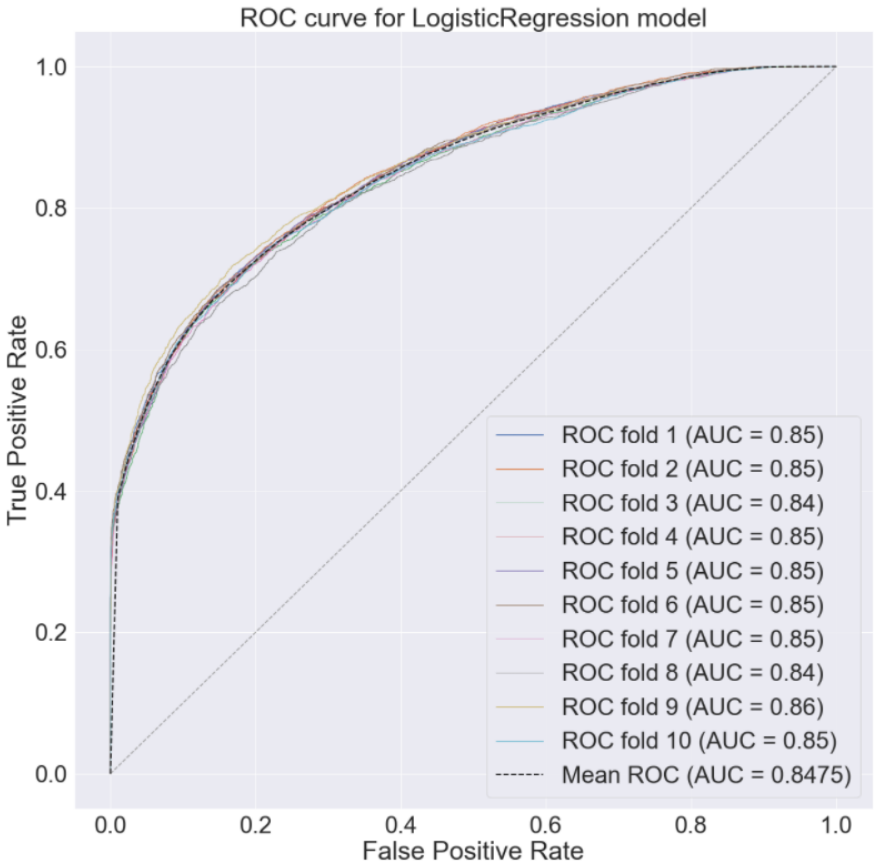


2.6 **בדיקת שונות מוסברת בעזרת PCA:**

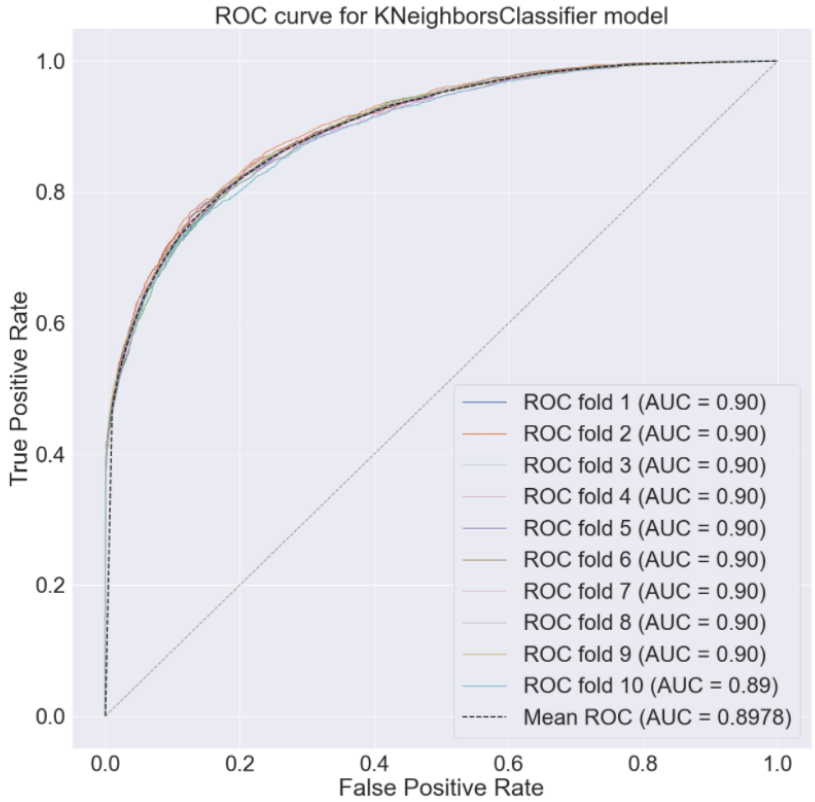


3. נספחי הרצת מודלים:

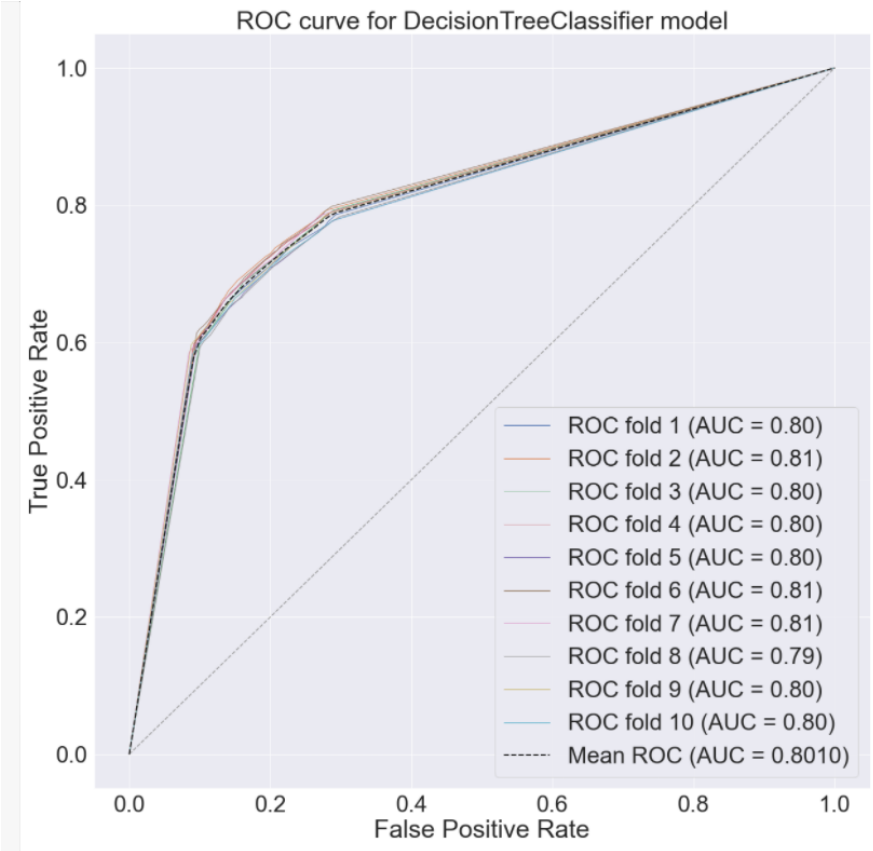
3.1 - Logistic Regression



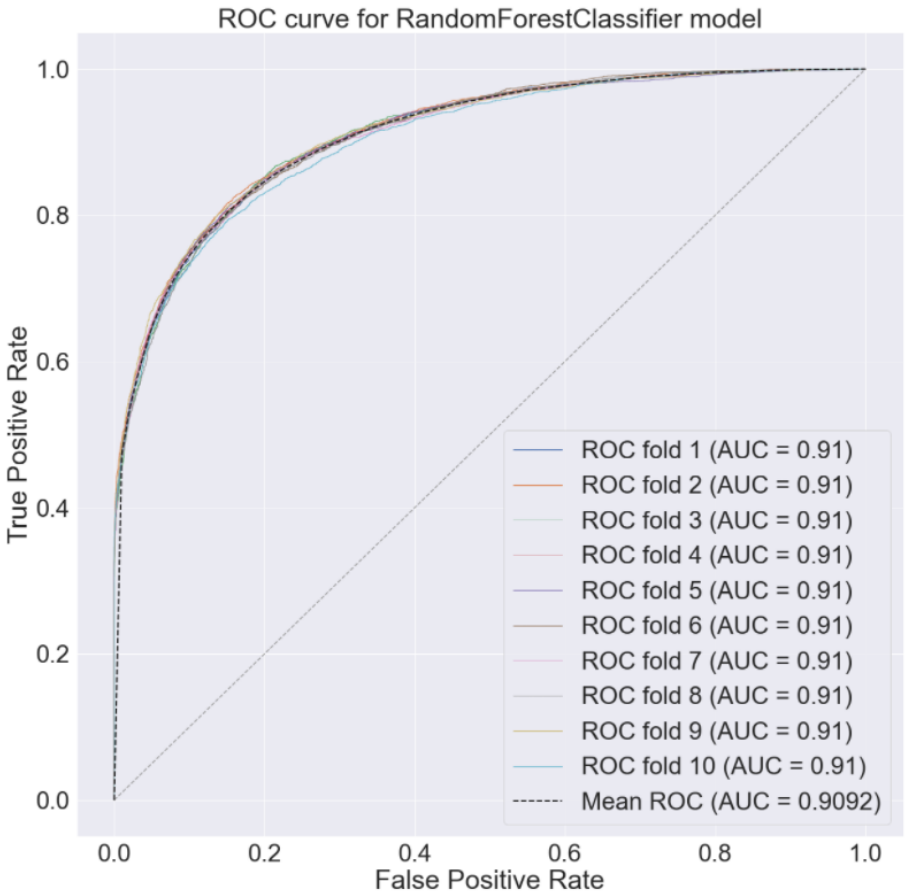
3.2 KNN



3.3 - Decision Tree



3.4 - Random Forest



3.5 - Confusion Matrix