פרויקט גמר בינה עסקית

מגישים: יעל אברהם 206656589 , עדן פבריקנט 212302541 ,

יאמן סלאלחה 207391236 .

הקדמה והנושא שבחרנו

אנחנו בחרנו בלעשות את העבודה על הזמנות במלון. ה Dataset נלקח מ Kaggle. הוא מכיל 120 אלף רשומות ו32 עמודות עם פרטים שונים.

ל Dataset קוראים- "Hotel booking demand"- הDataset מתאר הזמנות של חדרים ב2 מלונות בפורטוגל (מלון עירוני ומלון נופש).

קישור לDataset - https://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand?select=hotel_bookings.csv

השאלה המחקרית

מהם הגורמים המרכזיים שמשפיעים על ביטול הזמנות במלון, כיצד ניתן לחזות מראש הזמנות בסיכון לביטול, וכיצד ניתן להשתמש במידע זה כדי לשפר את שימור הלקוחות, למשוך לקוחות רווחיים יותר, ולהגדיל את הרווח הכולל של המלון?

מטרות

1. **חיזוי ביטולים מראש**

באמצעות ניתוח מאפיינים כמו: מדינה, ערוץ הזמנה, סוג חדר, לקוח חוזר, תאריך ההזמנה, משך השהות ועוד.

2. **זיהוי פרופיל הלקוח הנאמן**

מי הלקוחות שלא מבטלים, מבצעים הזמנות חוזרות, מזמינים חניה ובקשות מיוחדות - ניתן להתאים להם הצעות אישיות.

3. **פילוח לקוחות לפי ערך**

לקוחות רווחיים (לעומת כאלה שמובילים לביטולים/הפסדים) למשל כאלה עם adr נמוך וביטול גבוה.

4. **שיפור תהליכי שיווק ומכירה**

התאמת קמפיינים פר ערוץ הזמנה או סוג שוק - (market_segment) ניתוח איזה ערוץ מביא לקוחות טובים יותר.

5. **בניית אסטרטגיות שימור**

לדוגמה: לשלוח תזכורת לפני ההגעה ללקוחות בסיכון גבוה לביטול, או להציע שדרוג ללקוח חוזר.

תהליך ETL

את תהליך הETL ביצענו ב python בקובץ בשם cleaner.py. תחילה מחקנו עמודות לא רלוונטיות והשלמנו ערכים חסרים.

שינינו את הטיפוס של העמודות בהתאם וייצאנו את הקובץ הנקי.

תהליך DWH

חילקנו את הדאטה נקייה ל7 ממדים וטבלת עובדות אחת על מנת ליצור סכמת כוכב.

נסביר בקצרה כל טבלה -

Fact_Bookings - טבלת עובדה - מכילה את פרטי ההזמנות וקישוריות לשאר הטבלאות

Dim_Hotel - מכילה את סוג המלון שאליו בוצעו ההזמנות

Dim_Customer - מכיל מידע על הלקוח: מדינתו, האם לקוח נאמן

Dim_Date - עדכון סטטוס אחרון של ההזמנה

Dim_Agent - סוכני ההזמנות כאשר 0 מייצג הזמנה ישירה דרך המלון

Dim_MarketSegment - מתארת את איך ההזמנה בוצעה, דרך מי וסוג הלקוח

Dim_Meal - סוג הפנסיון

Dim_RoomType - סוג החדר שנבחר ע"י הלקוח וסוג החדר שניתן בפועל

Dim_Order_Status – סוג סטטוס הזמנה

ניתוחים עסקיים - דוחות BI ותובנות עיקריות

בחלק זה בוצעו ניתוחים עסקיים באמצעות Tableau במטרה להבין את הגורמים המשפיעים על ביטולים והכנסות במלון. להלן תמצית הדוחות העיקריים (9 דוחות) , התובנות וההמלצות העסקיות.

דוח 1: שיעור ביטולים לפי מדינה

מטרת הניתוח:
לזהות מדינות עם שיעור ביטולים גבוה לצורך התאמת מדיניות ופעילות שיווקית.

תובנות עיקריות:

- שיעורי ביטולים חריגים במיוחד באיחוד האמירויות (84.3%), ערב הסעודית (68.7%) ואינדונזיה.(68.5%)

המלצות עסקיות:

- החמרת מדיניות הביטולים באזורים אלה.
- שיפור התקשורת השיווקית במדינות עם שיעורי ביטול גבוהים.

דוח 2: ניתוח ביטולים לפי סגמנט וערוץ הפצה

מטרת הניתוח:
איתור סגמנטים וערוצים עם שיעורי ביטול חריגים לצורך שיפור יציבות ההכנסות.

תובנות עיקריות:

- פלח Groups מוביל עם 61% ביטולים, TA/TO עם 41.03%.
- שיעור ביטול נמוך ויציבות גבוהה בפלחי Direct ו Corporate.

המלצות עסקיות:

- להחמיר את תנאי ההזמנה מול קבוצות וערוצי - TA/TO למשל, דרישת מקדמה או קנסות ביטול.
- תעדוף ערוצי הפצה יציבים כמו Direct ו Corporate בקמפיינים עתידיים ובהטבות.
- לעקוב אחרי סוכנים וקבוצות עם הרבה ביטולים ולדרג אותם לפי רמת סיכון כדי לנהל אותם בצורה טובה יותר.

דוח 3: הצלבה בין סוג חדר מוזמן למוקצה

מטרת הניתוח:
בדיקת התאמה בין חדרים מוזמנים לחדרים שסופקו בפועל.

תובנות עיקריות:

- כ־7,548 הזמנות שונו מחדר A לחדר D.
- רוב ההזמנות (73,598) מסוג A סופקו כראוי.

המלצות עסקיות:

- לבדוק את הסיבה לשינויי סוג החדר - האם השינויים נגרמים בגלל מחסור בזמינות של חדרים מסוג ,A ואם כן - להגדיל את זמינות חדרים אלה..
- להקפיד על הודעה מראש ללקוחות בכל פעם שמתבצע שינוי.
- לבדוק כיצד השינוי השפיע על שביעות הרצון של הלקוחות ועל שיעורי הביטולים.

דוח 4: שיעור ביטולים לפי זמן מראש

מטרת הניתוח:
לבחון השפעת מועד ההזמנה מראש על שיעורי ביטול.

תובנות עיקריות:

- ביטולים נמוכים בטווח קצר (18.5%), גבוהים בטווח רחוק (67.6%).
- ככל שזמן ההמתנה מראש ארוך יותר, כך שיעור הביטולים עולה.

המלצות עסקיות:

- לקבוע מדיניות ביטולים מחמירה או תמחור מיוחד להזמנות ארוכות טווח.
- חיזוק הקשר עם הלקוח ככל שמתקרב מועד השהייה.

דוח 5: שיעור ביטולים לפי סוג ארוחה

מטרת הניתוח:
א לבדוק האם סוג הארוחה המוזמן משפיע על שיעור הביטולים, ולזהות את סוגי הארוחות שדורשים שינוי בתמחור או שיווק מיוחד.

תובנות עיקריות:

- FB עם שיעור ביטול גבוה במיוחד (59.9%) . למרות שמספר ההזמנות יחסית קטן (798 בלבד)
- BB עם שיעור מתון יותר (37.3%) עם 92,310 הזמנות.
- Undefined עם שיעור הביטול הנמוך ביותר - 24.5%.

המלצות עסקיות:

- לבדוק את המחיר ותנאי ההזמנה של חבילת FB ולשקול הנחות או תנאים טובים יותר ללקוחות פוטנציאליים.
- לבחון אם כדאי לצמצם את שיווק FB ולהגדיל את השיווק של חבילות BB או SC המשתלמות יותר למלון.
- למקד את השיווק של חבילת FB לקהלי יעד מתאימים, שיוכלו להגדיל את הסיכוי לשהייה בפועל.

דוח 6: שיעור ביטולים לפי סוג פיקדון

מטרת הניתוח:
בדיקת השפעת סוג הפיקדון על שיעור הביטולים.

תובנות עיקריות:

- ללא החזר עם שיעור ביטולים חריג (99.36%) – ייתכן שיש בעיית תקשורת או הבנה של תנאי הפיקדון.
- הזמנות ללא פיקדון מציגות ביטול נמוך יחסית (28.38%) עם מספר ההזמנות הגבוה ביותר (104,641).
- הזמנות עם החזר כספי כמעט ואינן קיימות (162 בלבד) אך מציגות שיעור ביטול של 22%.

המלצות עסקיות:

- שיפור הצגת והבהרת תנאי הפיקדון.
- תהליך אישור כפול להזמנות של אי החזר.
- חשוב לציין שרוב ההזמנות (104,641) היו ללא פיקדון, לעומת הזמנות של אין החזר כספי (Non Refund) - כולל רק 14,587 הזמנות. למרות המספר הקטן יחסית, סוג האין החזר כספי (Non Refund) מייצג בעיה חמורה עם כמעט 100% ביטולים, שדורשת בדיקה דחופה.

דוח 7: השפעת נאמנות לקוחות על ביטולים

מטרת הניתוח:
לבדוק השפעת נאמנות לקוחות על שיעורי הביטול.

תובנות עיקריות:

- לקוחות חדשים מבטלים ב-37.8% , לקוחות חוזרים מבטלים רק ב-14.5% - עשוי להעיד על נאמנות גבוהה יותר מצד לקוחות חוזרים, היכרות עם תהליך ההזמנה, או תחושת ביטחון גבוהה יותר בשירות מצד הלקוח.

המלצות עסקיות:

- השקעה בשימור לקוחות קיימים דרך הטבות ומועדוני לקוחות.

דוח 8: טופ 10 סוכנים לפי שיעור ביטולים

מטרת הניתוח:
איתור סוכנים בעייתיים ואפקטיביים לניהול משופר של הזמנות.

תובנות עיקריות:

- סוכן 1 עם שיעורי ביטול חריגים (73.4%), סוכנים 9 ו240 עם שיעורי ביטולים גבוהים, סוכן 28 מצטיין (6.6%) .

המלצות עסקיות:

- מומלץ לבצע בדיקה ממוקדת מול סוכנים בעייתיים כגון סוכן 1 (שיעור הביטולים הגבוה ביותר - 73.4%) וסוכן 9 (שני בשיעור הביטולים והסוכן עם מספר ההזמנות הגבוה ביותר).
- יש לבחון כיצד ניתן להגביר את פעילותו של סוכן 28, שמציג ביצועים מצוינים (רק 6.6% ביטולים), אך עם מספר הזמנות נמוך משמעותית.

דוח 9: ממוצע הכנסה לפי סגמנט

מטרת הניתוח:
איתור פלחי שוק רווחיים והפסדיים לצורך שיפור רווחיות המלון.

תובנות עיקריות:

- פלחי השוק הרווחיים ביותר הם Online TA (117.2) ו-Direct (115.4).
- פלחי השוק הפחות רווחיים הם Complementary (2.9) ו-Undefined (15).
- ישנו פער משמעותי בהכנסה בין פלחי שוק אלו, שמעיד על פוטנציאל שיפור ההכנסות באמצעות פילוח שוק מדויק יותר.

המלצות עסקיות:

- השקעה מוגברת בפלחי השוק הרווחיים.
- לשקול להגביל או לשנות את המדיניות עבור פלחי שוק הפסדיים.

<u>כריית מידע -Data mining:</u>

<u>ניתוח - Logistic Regression :</u>

מטרה:

המטרה של ניתוח זה הייתה ליישם מודל Logistic Regression לחיזוי האם הזמנה תבוטל (is_canceled) על בסיס מאפיינים שונים מתוך נתוני ההזמנות.

נתונים ומאפיינים:

בקובץ Fact_Booking (1).csv קיימים הנתונים הבאים ששימשו אותנו לבניית המודל:

- זמן מראש להזמנה (lead_time)
- לילות סוף שבוע (stays_in_weekend_nights)
- לילות אמצע שבוע (stays_in_week_nights)
- מספר מבוגרים (adults)
- מספר ילדים (children)
- מספר תינוקות (babies)
- האם מדובר באורח חוזר (is_repeated_guest)
- מספר ביטולים קודמים (previous_cancellations)

- מספר הזמנות קודמות שלא בוטלו (previous_bookings_not_canceled)
- סוג פיקדון (deposit_type)
- סוג לקוח (customer_type)
- מחיר ממוצע ללילה (adr)
- מספר בקשות מיוחדות (total_of_special_requests)
- מספר שינויים בהזמנה (booking_changes)

מתודולוגיה:

1. **עיבוד נתונים:**
   - הסרת ערכים חסרים.
   - המרת משתנים קטגוריאליים (deposit_type, customer_type) לערכים מספריים באמצעות LabelEncoder .
   - המרה של עמודת המטרה Is_canceled לערכים בינאריים (0- לא בוטל ,1בוטל)
2. **חלוקת הנתונים:**
   - הנתונים חולקו למערכי אימון ובדיקה ביחס של 80/20.
3. **בניית המודל:**
   - יושם מודל Logistic Regression על הנתונים.
4. **מדדי הערכה:**
   - Accuracy, Precision, Recall, F1 Score ו-ROC Curve.

תוצאות:

- Accuracy: 0.7737
- Precision: 0.8938
- Recall: 0.4512
- F1 Score: 0.5997

```
Accuracy: 0.7737
Precision: 0.8938
Recall: 0.4512
F1 Score: 0.5997
```

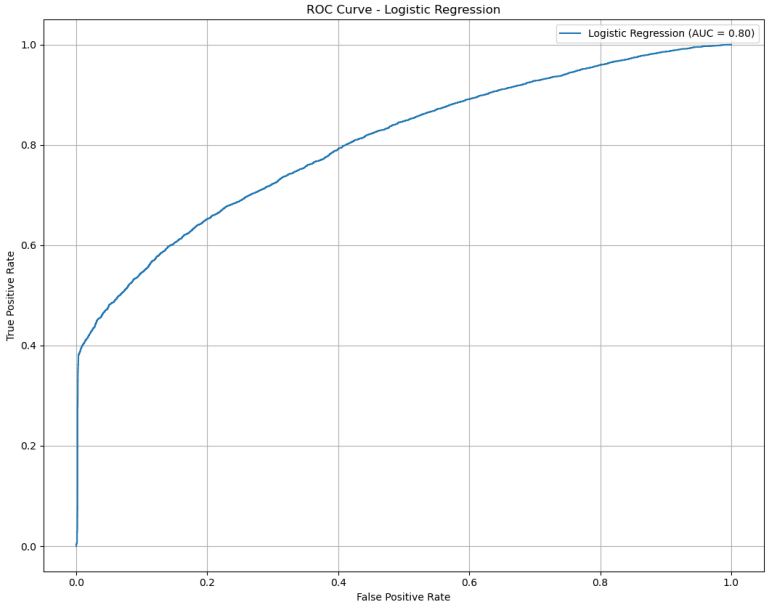
ה ROC Curve - מציג את הביצועים של המודל על ידי השוואת שני פרמטרים:

- **ציר ה X:**-שיעור התראות שגויות - (False Positive Rate) ככל שהערך קרוב ל 0, המודל פחות טועה בזיהוי שליליים כחיוביים.

- **ציר ה :Y-**שיעור זיהויים נכונים (True Positive Rate) - ככל שהערך קרוב ל-1, המודל מזהה טוב יותר את החיוביים.

AUC (שטח מתחת לעקומה):

- ערך ה AUC הוא 0.80 - זה אומר שהמודל מצליח לזהות הזמנות שבוטלו ברמת דיוק טובה יחסית (80%) .
- ככל שהעקומה קרובה לפינה השמאלית העליונה, כך המודל טוב יותר.



מסקנות והמלצות:

- המודל Logistic Regression הציג דיוק סביר (77.37%) , אך Recall נמוך (45.12%), מה שמעיד על קושי בזיהוי ביטולים בפועל.
- כדי לשפר את ביצועי המודל, מומלץ לבחון את האפשרות להוסיף מאפיינים נוספים שעשויים להשפיע על ביטול ההזמנה, לבצע כיול פרמטרים (Hyperparameter Tuning) ולהתמקד בהגדלת ה -Recall על חשבון ה -Precision במידה וחשוב יותר לזהות ביטולים בפועל.

<u>השוואת מודלים מונחי מטרה - חיזוי סטטוס ההזמנה</u>

מטרה:

המטרה של ניתוח זה הייתה ליישם מודלDecision Tree מודלים לחיזוי סטטוס ההזמנה (Check-Out, Canceled, No-Show) עבור הזמנות במלון על בסיס מגוון מאפיינים.

נתונים ומאפיינים:

השתמשנו בקובץ הנתונים Fact_Booking (1).csv, הכולל מידע מפורט על הזמנות במלון, תוך שימוש במאפיינים הבאים:

- זמן מראש להזמנה(lead_time)
- לילות סוף שבוע(stays_in_weekend_nights)
- לילות אמצע שבוע(stays_in_week_nights)
- מספר מבוגרים(adults)
- מספר ילדים(children)
- מספר תינוקות(babies)
- האם מדובר באורח חוזר(is_repeated_guest)
- מספר ביטולים קודמים(previous_cancellations)
- מספר הזמנות קודמות שלא בוטלו(previous_bookings_not_canceled)
- סוג פיקדון(deposit_type)
- סוג לקוח(customer_type)
- מחיר ממוצע ללילה(adr)
- מספר בקשות מיוחדות(total_of_special_requests)
- מספר שינויים בהזמנה(booking_changes)

מתודולוגיה:

1. **עיבוד נתונים:**
   - הסרת ערכים חסרים לשמירה על שלמות הנתונים.
   - המרת משתנים קטגוריאליים (deposit_type, customer_type) לערכים מספריים באמצעות LabelEncoder.
2. **חלוקת הנתונים:**
   - הנתונים חולקו למערכי אימון ובדיקה ביחס של 80/20 כדי להעריך את המודל.
3. **בניית מודלים:**
   - יושמו שלושה מודלים:
     - Decision Tree Classifier
     - Random Forest Classifier
     - K-Nearest Neighbors (KNN)

4. **מדדי הערכה:**

- חישוב דיוק Accuracy, Precision, Recall ו F-1 Score עבור כל מודל.
- יצירת גרף ROC Curve עבור Decision Tree להמחשת היחס בין True Positive Rate ל - False Positive Rate.

תוצאות:

תוצאות מודל: Decision Tree :

Accuracy : 0.7674 , Precision: 0.8280 , Recall:0.7675 , F1 Score : 0.7320

תוצאות מודל: Random Forest :

Accuracy : 0.8285 , Precision: 0.8251 , Recall:0.8285 , F1 Score : 0.8242

תוצאות מודל: KNN :

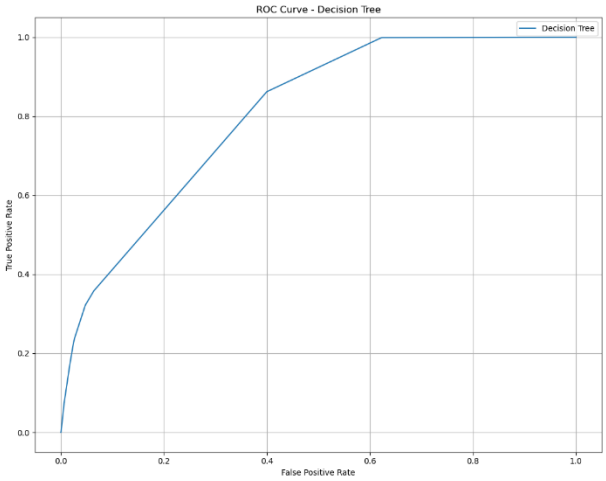
Accuracy : 0.7678 , Precision: 0.7640 , Recall:0.7678 , F1 Score : 0.7619

```
Decision Tree - Accuracy: 0.7675, Precision: 0.8280, Recall: 0.7675, F1 Score: 0.7320
Random Forest - Accuracy: 0.8285, Precision: 0.8251, Recall: 0.8285, F1 Score: 0.8242
KNN - Accuracy: 0.7678, Precision: 0.7640, Recall: 0.7678, F1 Score: 0.7619
```

ניתוח חשיבות מאפיינים: (Decision Tree)

המאפיינים המשפיעים ביותר על החלטות המודל:

- deposit_type – 0.673
- lead_time – 0.104
- previous_cancellations – 0.090
- adr – 0.050
- total_of_special_requests – 0.044



ROC Curve - Decision Tree :

ה ROC Curve- מציג את הביצועים של מודל Decision Tree ביחס לזיהוי ביטולים (is_canceled) .

- **ציר ה -X:** שיעור הטעויות החיוביות - (False Positive Rate) - ככל שהערך נמוך יותר, המודל טוב יותר במניעת טעויות חיוביות.

- **ציר ה -Y :**שיעור הזיהויים הנכונים - (True Positive Rate) ככל שהערך גבוה יותר, המודל מצליח לזהות ביטולים בצורה טובה יותר.

פירוש הגרף:

- העקומה קרובה יותר לאלכסון מאשר לפינה השמאלית העליונה, מה שמראה על ביצועים בינוניים.
- במודל ,Decision Tree ה - AUC(שטח מתחת לעקומה) אינו גבוה במיוחד, מה שמעיד על כך שהמודל לא מצליח להבחין בצורה טובה בין ביטולים להזמנות שלא בוטלו.

מסקנות והמלצות:

- מודל Random Forest הראה את הביצועים הטובים ביותר עם דיוק ו- Recall גבוהים יותר.
- deposit_type היה המאפיין המשמעותי ביותר בקביעת סטטוס ההזמנה, מה שמעיד על קשר חזק בין סוג הפיקדון לתוצאה.
- ניתן לשפר את המודלים על ידי כיול פרמטרים נוספים וביצוע הנדסת מאפיינים.

<u>ניתוח Clustering:</u>

מטרה:

ביצענו ניתוח Clustering במטרה לזהות קבוצות לקוחות עם מאפייני ביטול דומים בהתבסס על שני משתנים מרכזיים:

1. Lead Time - זמן מראש להזמנה.
2. Previous Cancellations-מספר הביטולים הקודמים של הלקוח.
המטרה הייתה לזהות קבוצות לקוחות עם דפוסי התנהגות דומים ולהבין מי הם הלקוחות בעלי הסיכון הגבוה לביטול.

נתונים ומאפיינים:

השתמשנו במאפיינים הבאים מתוך הקובץ : Fact_Booking (1).csv

- זמן מראש להזמנה (lead_time)
- מספר ביטולים קודמים (previous_cancellations)

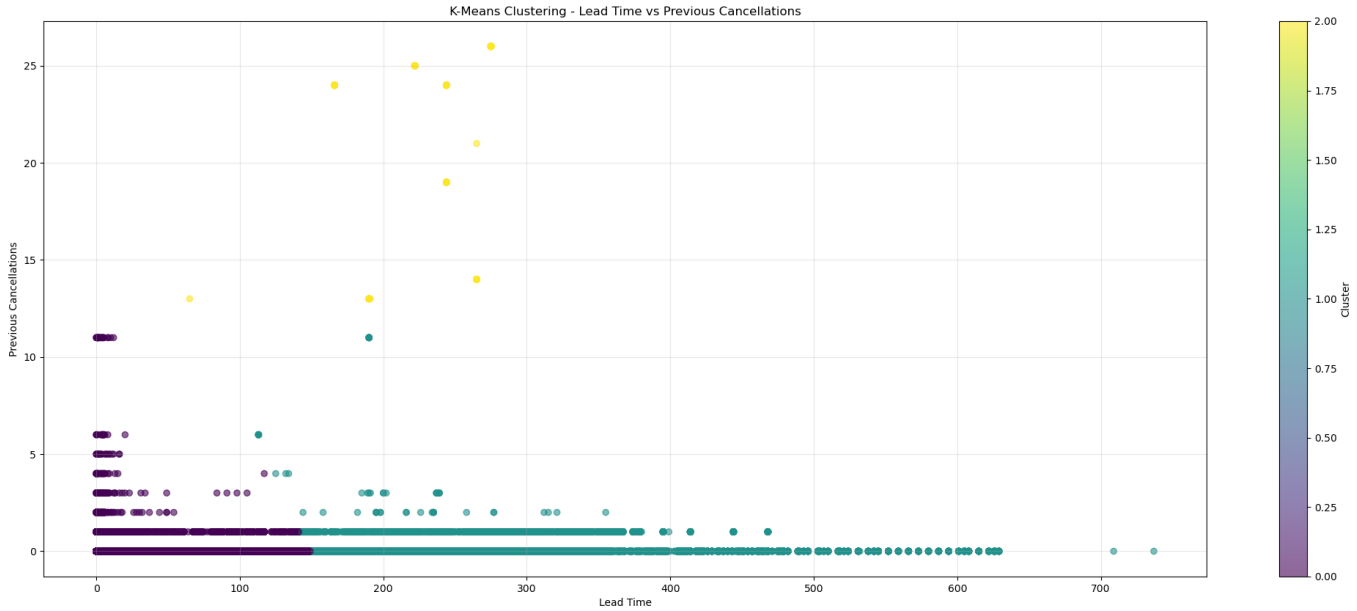
מתודולוגיה:

1. **עיבוד נתונים:**
   - הסרת ערכים חסרים להבטחת שלמות הנתונים.

- נרמול הנתונים באמצעות StandardScaler() על מנת למנוע הטיות שנובעות מערכים גדולים במיוחד.

2. ביצוע אשכולות (Clustering):

- יישום K-Means Clustering עם 3 אשכולות.



תוצאות: K-Means Clustering

בגרף ה K-Means - ניתן לראות את שלושת האשכולות שסומנו בצבעים שונים:

טבלה מסכמת של מאפייני האשכולות:

אשכול	מספר לקוחות בקבוצה	ממוצע Lead Time	ממוצע Previous Cancellations
0	23,384	40.2	0.1
1	10,932	82.5	1.3
2	1,250	69.8	8.9

- אשכול 0 (סגול) - לקוחות יציבים עם ביטולים נמוכים וזמן המתנה קצר.
- אשכול 1 (טורקיז) - לקוחות עם ביטולים מתונים וזמן הזמנה בינוני.
- אשכול 2 (צהוב) - לקוחות עם ביטולים רבים וסיכון רב.

ניתוח חריגות:

- זוהו מספר נקודות חריגות עם מספר ביטולים גבוה מאוד (מעל 15 ביטולים) וזמן הזמנה קצר.

- זוהו מספר נקודות חריגות עם ביטולים רבים בזמן הזמנה קצר.

מסקנות:

זוהו שלוש קבוצות עם רמות סיכון שונות: נמוך (הזמנות קצרות), בינוני (משולב), וגבוה (ביטולים רבים).

המלצות לפי אשכול :

<u>אשכול 0 – לקוחות יציבים עם סיכון נמוך לביטול:</u>

- מומלץ להשקיע בשימור לקוחות אלו באמצעות הצעות לשדרוג והטבות לשימור.
- מדובר בקהל אמין שניתן לבנות עליו לטווח הארוך.

<u>אשכול 1 - לקוחות בעלי סיכון בינוני:</u>

- יש לעקוב אחר התנהגותם ולבצע קמפיינים שיווקיים או תמריצים להזמנה מוקדמת.
- מומלץ להפעיל תזכורות ותזכורות לקראת מועד השהייה.

<u>אשכול 2 – לקוחות בעלי סיכון גבוה מאוד לביטול:</u>

- מומלץ להטמיע מנגנון אישור כפול או דרישת מקדמה להזמנות חדשות מקבוצה זו.
- יש לשקול סינון מוקדם של לקוחות עם היסטוריה בעייתית או הגבלת אפשרות ביטול.
- ניתן לבדוק התאמה להצעות שונות - לדוגמה: רק תשלום מראש, או רק חדרים בסיסיים.

תובנה מערכתית:
שילוב בין קלאסטרינג למודלים חיזויים מאפשר הבנת קבוצות ברמה אסטרטגית וחיזוי פרטני, וכך ניתן לפעול באופן מדויק מול כל לקוח וקבוצה.

כך ניתן לתכנן קמפיינים שיווקיים, הטבות, או דרישות תשלום שונות, לפי גם רמת הסיכון של הלקוח וגם מאפייני הקבוצה שאליה הוא שייך.