



המחלקה להנדסת תעשייה וניהול

שם הפרויקט: מערכת לניהול לידים באופן חכם Leadest

ספר הפרויקט

<u>דניאל לבקוביץ'</u> , טניה פילוחוף	שם הסטודנט:
<u>אבי עסיאו</u>	שם המנחה:
<u>אבי עסיאו</u>	מציע הרעיון:
<u>29.5.2022</u>	תאריך ההגשה:

אישור של המנהה להגשת דוח הסופי

<Avi Assis <aviassis10@gmail.com

אל: Daniel Levkovitz

עותק: Tanya Filozof

AA

דניאל וטניה,

לאחר שעכברתי על מסמכי הפרויקט - בהחלטת עבודה מקצועית וטובה ומעמיקה,
מאשר הגשת הפרויקט. יש להגיש גם טופס קיטלוג בהתאם לסטנדרטים (קובץ אקסל).

הרבה הצלחה,

אבי עסיס

...

תודות

ברצוננו להודות למנהל שלנו, מר אבי עסינס, אשר נתן לנו יד חופשית להגשים את החזון שלנו במערכת, משלב האפיון ועד שלב ביצוע הפרויקט.

תקציר מנהלים

כיום, יש המונ ארגונים מתעשית הרכב והלייניג (חברות פרטיות ועד ארגונים גדולים), שאופן ביצוע מכירות ליסינג לרכבים מתבצעת בשיטה מיושנת.

שיטה מיושנת משמעותה שכאשר אדם או חברה מתעניינים בשירות של השכרת רכב - הם משאירים את הפרטים שלהם באתר החברה, טלפון, רשות חברות וכו', ולאחר מכן אכן נשיכת השוים חוזרים בשיחת טלפון לאותם מתעניינים בשיטת פיפ'ו (First In First Out) שבו בפועל את המכירה.

בפרויקט זה הוקמה מערכת המלצות חכמה אשר מדרגת את הלידים בשבייל להגדיל את הסיכוי לממכר ותאפשר את שיטת העבודה של אנשי המכירות בעת השימוש בה. לצד מילון ודרוג הלידים, המערכת מציגה Dashboard המכיל גרפים המסייעים בניתוח הלידים והמכירות בכך שמנaily הסניף יוכל לקבל תמונה מלאה של המכירות ומה הם הפרמטרים החשובים ביותר של ליד בצד' שיופיעו לממכר.

אחד היתרונות הבולטים בפרויקט הוא שהמליצה נעשית על ידי שני סוגי שוניים של אלגוריתמי למידת מכונה, למידה מונחית ולמידה בלתי מונחית. לאחר ההרשמה לאתר, המשמש מעלה קובץ לידים, קובץ נתונים אשר מכיל מידע על לקוחות פונטיציאליים אשר הביעו התעניינות לגבי חוזה ליסינג.

אותו קובץ עובר תהליכי של הכנת הדטה: כמו סינון, וטרנספורמציה של הנתונים. בסיום התהליך הראשוני, אלגוריתם משפחתי הלמידה הבלתי מונחית מופעל ומבצע חלוקה ראשונית לארבעה קבוצות: ליד "רותח", "חם", "בינוני" ו"קר".

בפעם השנייה שהמשתמש עושה במערכת שימוש, לאחר שחברת הליסינג השתמשה בקובץ המקורי וניסתה לבצע מכירה, מופעלים אלגוריתמיים משפחתי הלמידה המונחית אשר לומדים אילו לידים הפכו למכירה. לבסוף, האלגוריתם בעל אחוז הדיוק הטוב ביותר ביוטר יישמר בענן המשמש בשלב השלישי והאחרון, המערכת מסוגת על פי המודל הטוב ביותר שנשמר בענן, וחוזה אילו לידים בעלי המאפיינים המתאימים ביותר יהפכו למכירה ממשית. בכך שהמערכת תהיה אינטראקטיבית, המשמש מקלט בכל שלב מיליל כתובות המייל שהזונה במהלך ההרשמה והקבצים נשמרים בענן "יעדי" של המשתמש וזאת באמצעות שימוש ב-API's Google's.

מסמרק זה מסכם את פרויקט הגמר ומכל שכלל של מסמרק-WoS ודוח הביניים שנכתבו במהלך בניית המערכת והכילו את האפיון, התכנון והגדרות של המערכת.

לאחר סקירת המצב הנוכחי בארץ ו בחו"ל, תוך הגדרת בעלי העניין, הצלחנו לאfine את מטרת, יעדיו ומדדי הפרויקט ובחרנו בחולופה המועדף להקמת המערכת - פיתוח עצמאי תוך שימוש בחבילות חינמיות בתוכנות במשפט הקוד Python, תוך התאמה לדרישות הפונקציונליות, בזמן ולעולות הנדרשים לפרויקט. סקירת הספרות אפשרה להעמק בחולופות המגוונות בעולם התוכנות על מנת לבנות את המערכת והאתר והסקנו כי באמצעות שימוש ב-*Python*, תוך כדי שילוב של תוכנות מונחה עצמים נוכל לבנות מערכות דינמיות עבור המשתמש. כמו כן, שפת הפיתוח של בניית האלגוריתמים השונים הינה Python מהווה שחקן ראשי בעולם ה-Data Science וכוללת בתוכה אפשרות רבות של למידת מכונה.

בסיס הנתונים שנעשה בו שימוש הוא SQLite, שפות פיתוח הצד לקוח הן בשילוב של Python, HTML, CSS, Javascript, Flask, Bootstrap. במערכת זו נעשה שימוש בקובץ לידים מוגנה לשימוש יצרתו, השתמשנו ב-Data Generator על ידי שימוש בתנאים המדים את עלים הליסינג האמתי.

פיתוחים עתידיים אשר אנו צופים למערכת הינה שיפור האלגוריתם המונחה על ידי שימוש בקובץ נתונים יותר מואזן על מנת לשפר את תוצאות הדיוק אשר הגיעו לדין של 83% אחוז במהלך הבדיקות שלנו.

מסמרק זה מפרט את תיקון המערכת באמצעות הסברים מילוליים, תרשימים, טבלאות והתהליכי של המערכת.

Executive Summary

Most organizations from the automotive and leasing industry (private companies to large organizations) make the sale of car rentals done in an outdated method.

An outdated method means that when a person or company are interested in a car rental service - they fill their details on the leasing company website, Company's Social Networks, or contact the leasing company by phone. Then, the various salespeople contact those potential customers in the FIFO method (First In- First Out) to convert the lead to a sale.

As a result of recognizing this uncomfortable and inefficient method leasing companies use, a recommendation system has been developed in this project. With the recommendation system, the aim is to rank the leads and define those with the highest probability of converting into actual sales while simplifying the salesperson's job and increasing the company's profit.

In addition, the system displays a dynamic dashboard that contains graphs that help analyze the leads and sales so that branch managers can have more meaningful insights into the leads' distribution and characterize the most significant parameters of the ones that are turning into actual sales.

Our project has the advantage of utilizing two different types of machine learning methods: supervised and unsupervised learning.

After registering on the site, the user uploads a leads file. The data file received from the user undergoes several processes: data preparation, filtering, and data processing. At the end of the process, the algorithm from the unsupervised learning is activated and divides the data into four groups: hot, high, medium, and low.

The second time the user uses the system, and the leasing company has used a group's division file to make a sale, algorithms from the supervised learning family are activated which learn which leads have converted to a sale. Finally, the algorithm with the highest accuracy will be maintained in the user's cloud.

In the third and final stage, the system classifies a file according to supervised learning. It predicts which leads have the most appropriate characteristics to be converted to an actual sale. For the system to be interactive, at each stage, the user receives an email to the email address documented during the registration, and the files are stored in the user's dedicated cloud while using Google's API.

This document summarizes the final project and contains a weighting of the SOW document and the midterm report written during the construction of the system which contained the characterization, planning, and definition of the system.

As a result of reviewing the existing worldwide situation, focusing on Israel, and defining stakeholders, we were able to determine the project's goals, objectives, and metrics. By defining all those metrics, we were able to decide on the best option for establishing the system - an independent development using free programming packages in Python code, while also adjusting the functional requirements, time, and budget for the project. The literature has made it possible to dive deeper into the diverse alternatives in the world of programming to build the system and site. We discovered different methods and noticed that Python is also

considered a leading actor in the Data Science world that obtains many machine learning capabilities.

Finally, we concluded that the combined tools of Python as a programming language for both developing the website and writing the algorithms while using the integration of object-oriented programming are the best solutions for a dynamic system that can provide all the user inquiries.

The database used is SQLite, and client-side development languages were combined with Python, HTML, CSS, and JavaScript along with the Bootstrap and Flask libraries. The system used a built-in lead file for its creation, we used the Data Generator by employing conditions that simulate the real world of leasing.

The future developments that we anticipate for the system are the improvement of the supervised learning algorithm which reached an accuracy of 83% during our tests.

This document details the system's design with documentation of verbal explanations, diagrams, tables, and the system's different processes.

תוכן עניינים

2.....	אישור של המנהה להגשת דוח הסופי
3.....	תעודות
4.....	תקציר מנהליים.....
5.....	Executive Summary
7.....	תוכן עניינים.....
10.....	רשימת איוורים טבלאות וגרפים
12.....	מילון מונחים.....
13.....	1 מבוא.....
14.....	2 מטרות יעדים ומדדים.....
14.....	2.1 מטרת הפרויקט.....
14.....	2.2 יעדי הפרויקט.....
14.....	2.3 מדי הצלחה.....
14.....	2.4 מדי כריית המידע / הפקת התובנות.....
15.....	2.5 עמידה במדדים.....
16.....	3 סקירת ספרות.....
16.....	3.1 מבוא לניהול לידים
17.....	3.2 טכנולוגיה ויישום
18.....	3.3 למידת מכונה – Machine Learning
19.....	3.4 תהליכי ה-Preprocessing
20.....	3.5 אלגוריתמי המערכת
22.....	3.6 מדי דיק.....
25.....	3.7 בסיסי נתונים - Databases
26.....	3.8 שירותים אחסון
26.....	3.9 שפות תכנות צד לקוח
27.....	3.10 שפות תכנות שכוללות כלים למטרת קלוסיפיקציה
28.....	4 מצב קיימ.....
28.....	4.1 סקירה ותיאור מצב קיימ
32.....	4.2 סקירת המצב הקיים בארץ ובעולם
33.....	5 מסמך דרישות
34.....	6 בחינה וניתוח חלופות מערכתיות
34.....	6.1 אלטרנטיבות לשימוש הפרויקט
39.....	6.2 תיקוף ובדיקות
53.....	7 אפיון המערכת
53.....	7.1 סיכום תהליכי כריית המידע, ETL וה-Pre-processing

61	7.2 אפיון המערכת – מערכות מידע
68	7.3 ניתוח חלופות טכנולוגיות
75	8 תיכון המערכת – System Design
75Network Diagram 8.1
76Component Diagram 8.2
77Class Diagram 8.3
78ERD – Entity Relationship Diagram 8.4
79	8.5 מיליון נתונים
90	8.6 דיאגרמת רצף או תרשימים פעילות לתהליכיים העיקריים במערכת
94	8.7 תרשימים עץ המסכים
95	9.1 תיאור של אלגוריתמים ותהליכי חישוב שמבצעת המערכת
95K - Means Clustering 9.1
96K - Prototypes Algorithm 9.2
97Logistic Regression 9.3
98Decision Tree 9.4
101Random Forest 9.5
103Grid Search CV 9.6
104	10.1 תוכרי הפרויקט: מערכות מידע
10410.1 מסכים של אתר האינטרנט
113Google Cloud Platform 10.2
11510.3 פלטי אמייל
117	11.1 תוכנות כריית המידע והמודלים שפותחו עבורו
11711.1 חלק הראשון – תוכנות כריית המידע של אלגוריתם K – Prototypes
12311.2 חלק השני – הלמידה המונחית
12911.3 חלק השלישי – חיזוי לדיים לפי המודל הטוב ביותר
131	12.1 בדיקות והערכת (System Testing and Evaluation)
13112.1 תכנית בדיקות מערכת (STP)
13112.2 תרחישי בדיקות (STD)
13112.3 דוח בדיקות מערכת (STR)
13112.4 הערכת המערכת המוצעת
133	13.1 תוכנית הפרויקט
13313.1 תוכנית עבודה. תרשימים גאנט מלא לכל מהלך הפרויקט
13413.2 תרשימים WBS
13513.3 ניהול סיכונים
137	14 סיום

137.....	14.1 סיכום ומסקנות
137.....	14.2 פיתוחים עתידיים והמשך עבודה
139.....	15 ריכוז שינויים מדויק התקן המפורט בפורמט טבלאי
140.....	16 רשימת מקורות.....
141.....	17 נספחים.....
141.....	17.1 מאמר באנגלית.....
141.....	17.2 תוכנית בדיקות המערכת - STP
142.....	17.3 מסמך עיצוב בדיקות המערכת STD – Software Test Description
144.....	17.4 דוח בדיקת אב טיפוס (STR)
145.....	17.5 פוטו
145.....	17.6 תיעוד ודף נתונים
145.....	17.7 מסמך ה-SOW המקורי
146.....	17.8 נספחים נוספים
146.....	18 אב טיפוס

רשימת איורים טבלאות וגרפים

מספר	שם	עמוד
.1	טבלה רשימת איורים טבלאות וגרפים	10-11
.2	טבלה מיליון מונחים	12
.3	תרשים 3.2.1 הדגמה של מערכת המלצה	17
.4	תרשים 3.1.1. התהילה שהמידע עובר בלמידה בלתי מונחת	18
.5	תרשים 3.3.2. דוגמה למערכת למידה מונחת	19
.6	תרשים 3.5.1 דוגמה של הפקנזה הסיגמאידית שמשמשת את הרגרסיה הלוגיסטיות	20
.7	תרשים 3.5.4. שינוי המרכזים של K-Means בכל איטרציה	22
.8	איור Confusion Matrix 3.6.2	23
.9	תרשים 3.6.7 תרשים שיטת המרפק	25
.10	טבלה 4.1 חממת החברות המובילות בשירותי הליסינג בישראל	28
.11	טבלה 4.1.4. תיאור סניף ליסינג	31
.12	טבלה 6.1.5. השוואה בין החלופות	37-38
.13	גרף 6.2.1.1.1.2. ניתוח הגברים מול הנשים	40
.14	גרף 6.2.1.1.1.3. ניתוח שעوت הפניה של הליד	40
.15	גרף 6.2.1.1.1.4. ניתוח המחלקה בה עבד הליד	41
.16	גרף 6.2.1.1.1.5. ניתוח הפלטפורמה ממנה הגיעו הלידיים	41
.17	גרף 6.2.1.1.1.6. חברות הרכבים הći מבוקשות על ידי הלידיים	42
.18	גרף 6.2.1.1.2.2. מחיר הרכב לעומת השנה הקודמת	43
.19	גרף 6.2.1.1.2.3. 15 חברות המכניות הcli שכיחות	43
.20	גרף 6.2.1.2.1. כמות המכירות שנסגרו	46
.21	גרף 6.2.1.2.2. כמות מכירות שנסגרו לעומת החולקה שביצעו בחלק הראשון	47
.22	גרף 6.2.1.2.3. שנת הרכב מול מחיר הרכב	47
.23	גרף 6.2.1.2.4. קבוצת הגילאים שהפכו למכירה	48
.24	גרף 6.2.1.2.5. התפלגות צפיפות המכירות גברים מול נשים	48
.25	גרף 6.2.1.3.1. ניתוח מכירות הגברים מול הנשים	49
.26	גרף 6.2.1.3.2. גילאי האנשים שהפכו למכירה לעומת כאלה שלא	49
.27	גרף 6.2.1.3.3. ניתוח שעת פניה הליד לעומת האם הוא נמכר או לא	50
.28	גרף 6.2.1.3.4. ניתוח מחלוקת שבת עובד הליד לעומת האם הוא הפרק למכירה או לא	50
.29	גרף 6.2.1.3.5. ניתוח פלטפורמות המכירה לעומת האם הם הפקו למכירה	51
.30	טבלה 7.2.1. בעלי עניין	61
.31	טבלה 7.2.2. שחיקנים	62
.32	תרשים 7.2.3. מקרי שימוש	63
.33	טבלה 7.2.4. מקרי שימוש	64-65
.34	טבלה 7.3.1.1. חלופות לשרת	68
.35	טבלה 7.3.1.2. פלטפורמה לבניית אתר	69

70	טבלה 7.3.1.3. מסד נתונים לאותר	.36
70	טבלה 7.3.1.4. פלטפורמת אחסון ענן	.37
71	טבלה 7.3.2.1. חלופות לאלגוריתמים הבלתי מונחים	.38
73	טבלת 7.3.2.2. אלגוריתמים משפחחת Supervised-Learning	.39
75	תרשים 8.1. Network Diagram	.40
76	תרשים 8.2. Component Diagram	.41
77	תרשים 8.3. דיאגרמת מחלקות	.42
78	תרשים 8.4. Erd	.43
79	טבלה 8.5.1. טבלת Users	.44
80	טבלה 8.5.2. טבלת Workers	.45
81	טבלה 8.5.3. טבלת LeadsFile	.46
82	טבלה 8.4. טבלת Branch	.47
83-84	טבלה 8.5.5.1. קובץ הלידים המתkeletal על ידי הילוקו	.48
85-86	טבלה 8.5.5.2. קובץ הנתונים לאחר שלב ה-ETL	.49
87	טבלה 8.5.5.3. מאגר המידע של הרכבים	.50
88-89	טבלה 8.5.5.4. מאגר המידע של החברות העסקיות	.51
90	תרשים 8.6.1. Activity Diagram – תהליך הרשמה לאותר	.52
91	תרשים 8.6.2. Activity Diagram – תהליך התחברות	.53
92	תרשים 8.6.3. Activity Diagram – תהליך צפיה בגרפים	.54
93	תרשים 8.6.4. Activity Diagram – תהליך העלאה קובץ	.55
94	תרשים 8.7. עץ המרכיבים	.56
96	נוסחה 9.2.1. פונקציית ההפסד עבור K-Prototypes	.57
98	טבלה 9.4.1. טבלת טרמינולוגיה עץ החלטה	.58
99	איור 9.4. מדגים את הטרמינולוגיה של עץ החלטה	.59
100	גרף 9.4.5. אנטropiphia	.60
100	איור 9.4.7. רוח מידע	.61
101	איור 9.5.1. מדגים את תהליך Bagging – שמבעץ העיר האקראי	.62
102	תרשים 9.5.2. דוגמה לביצוע והחלטה של יער אקראי	.63
102	טבלה 9.5.4. הבדלים בין עץ החלטה ל-Random Forest	.64
117	גרף 11.1.1. יישום שיטת המרפֶק	.65
119	טבלה 11.1.2.1. הסבר היפר פרמטרים של K – Prototypes	.66
125	טבלה 11.2.3. הסבר היפר פרמטרים של עץ ההחלטה	.67
133	תרשים 13.1. תרשימים גאנט	.68
134	תרשים 13.2. WBS	.69
135-136	טבלה 13.3. ניהול סיכונים	.70
139	טבלה 15. ריכוז שינוי	.71
142-143	טבלת 17.3. בדיקות STP	.72
144-145	טבלת 17.4. STR	.73

מילון מונחים

מונחים אשר חוזרים על עצם במהלך הפרוייקט רשומו במילון המונחים הבא.

מו'	מונח	הסבר
1.	לידים	לקוחות פוטנציאליים אשר הבינו התעניניות, או מיקוטלים ככאלה שהיו מעוניינים ברכישת מוצר או שירות.
2.	קובץ לידים	קובץ לידים זהו קובץ טקסט המכיל פרטיים על לקוחות שהבינו התעניניות על קניית מוצר או שירות.
3.	למידה בלתי מונחת	למידה בלתי מונחת היא מושג מלמידת מכונה. כאשר יש לנו מידע שהינו מעוניינים לקטalg / לחלק אותו לקבוצות (עוד מידע בסעיף הבא (3.3.1))
4.	למידה מונחת	למידה מונחת היא מושג מלמידת מכונה. כאשר 알고ירטם מסויים "לומד" קבוצה מסוימת ועל פי אותה למידה, הוא ידע לבצע סיווג. (עוד מידע בסעיף הבא (3.3.2))
5.	לייסינג	שיטת מימון לרכישת ציוד. בפרויקט שלנו המושג מתקשר להשכרת מכונית לטוויה רחוק (מעל 30 ימים).
6.	דאטה פריימ	טבלה המכילה מידע, אשר השורה הראשונה של אותה טבלה מתארת את המידע של שאר השורות.
7.	API	Application Program Interface. Uracha של ספריות קוד, פונקציות מוכנות, בהן יכולם המתכנתים לעשות שימוש פשוט, בלי להידרש לכתוב אותן בעצמם כדי שיוכלו להשתמש בטובות היישום שלהם.
8.	איטרציות	ריצה של קוד בתוך לולאה, כאשר קטע קוד רץ באופן איטרטיבי, כל מקטע קוד של אותה לולאה נקרא איטרציה.
9.	แดשborad	בעברית: לוח מחוונים. זהו מסך המציג מידע באופן גרפי אשר באמצעותו ניתן להסיק על המידע מסווג.

טבלה מילון מונחים

1 מבוא

במסגרת סקר השוק שביצעו בעת הגדרת צורף הפרויקט נחשפנו לתעשיית הרכב והלייסינג בימינו שmorכבת מחברות פרטיות וארגוני גדולים אשר מציעים שירותי מכירה והשכרה של רכבים בשיטות ארכאיות ולא ייעילות.

במהלך הקיימם היום אדם או חברה אשר מתעניינים בשירותי השכרת רכב באמצעות לייסינג - משאים פרטיים אישיים מאות פלטפורמות השיווק (אטר החברה, מודעת גוגל, פיסבוק וכו'), לאחר מכן אנשי המכירות השונים חוזרים אליום בשיחת טלפון על פי שיטת התזמנונים פיפ'ן (First In First Out).

במצב האופטימלי, אותו אנשי מכירות מספיקים לחזור אל כל לקוחות שהש亞וו פרטיים אר לעיתים קרובות מכירות רבות מתפספסות עקב לכך שאנשי המכירות מנוטים לבזבז את המשאבים שלהם (זמן העבודה) על ידיים שההסתברות שלהם להפוך למכירה היא נמוכה.

בנוסף, לחברות הליסינג הבינלאומיות בשוק, אין מערכות ניתוח סטטיסטיות ייעילות אשר מנתחות אילו ידיים הפכו למכירה ומה מאפיין אותם, מה שהיא יכולה לשפר את אסטרטגיית השיווק שלהם ובסופה של דבר להגדיל את המכירות של הסניפים.

לכן יזמונו להקים את פרויקט Leadest אשר נועד לשפר את העבודה היומיומית של אנשי המכירות והגדלת הרוחן של הארגון על ידי פיתוח מערכת המלצה ופיתוח אתר אשר תנהל את רשותות ה"ילדים" (מכירות פוטנציאליות) תמיין אותם לפי אחזוי הצלחת המכירות שלהם על ידי שימוש בכלים מעולים אנלטיקה העסקית.

בנוסף, תספק למנהל המכירות תמונה מצב של המכירות והילדים תוך כדי ניתוח של המאפיינים הבולטים שלהם.

הנתונים שיישמשו את המערכת יהיו נתונים עסקיים אשר יתקבלו על ידי המשתמש או נתונים נצברים של הארגון בקובץ טבלי (CSV), הקובץ זה יכיל "ילדים" כלומר ל��וחות פוטנציאליים שהבירו התעניינות לגבי המוצר של החברה, המערכת תשתמש בקובץ זהה בשבייל ליצור סיווג ומין ותחזיר אותם למשתמש בשבייל שיוכל להשתמש בהם לצורכי מכירה.

אנחנו נפתור את בעיות אלו באמצעות מערכת חכמה שתוכל לקבל המון מידע על לקוחות שונים לפי המאפיינים של הלקוח.

2 מטרות יעדים ומדדים

2.1 מטרת הפרויקט

בנייה מושלמת המאפשרת למשתמשי המכירות לנהל לידיהם פוטנציאליים בצורה יעילה ולחזות אילו לידים בעלי ההסתברות הגבוהה ביותר להפוך למכירה.

2.2 יעד הפרויקט

2.2.1 מקסום רוחן החברה לידיים קיימים.

2.2.2 ייעול זמני העבודה של אנשי המכירות - אנשי מכירות ישאפו ליצור קשר רק עם לידים עם אחוז הצלחה גבוהה.

2.2.3 הרחבת כושר המעקב וקבלת החלטות של מנהלי המכירות באמצעות יצירת דוחות כספיים על מצב הארגון באופן מייד.

2.2.4 יצירת "מסך משתמש גרפי" (GUI) בעל פונקציונליות פשוטה למשתמשי המערכת לצירוף סיווג לידים באופן מהיר.

2.2.5 יצירת התאמת אופטימלית של "לידים" (מכירות פוטנציאליות) לפי סיווג לאנשי המכירות העובדים בחברה.

2.3 מדדי הצלחה

2.3.1 הפיכת 30% מהילדים ל"לידים חמים" אשר יהפכו למכירה.

2.3.2 הגדלת רוחן הארגון ב-20%.

2.3.3 מזעור איבוד לידים פוטנציאליים ב 80% באמצעות שימוש במערכת התראות.

2.3.4 הסתגלות מהירה של משתמשי המערכת למערכת החדשה וסיווג לידים בצורה מהירה בכ-20% מהמערכת הקודמת.

2.4 מדדי כריות המידע / הפקת התוצאות

2.4.1 בשלב הראשון - קטלוג הילדים לפי 4 קבוצות איקוט: לידים "רותחים", לידים "חמים", לידים "בינוניים" ולידיים "חמים". המדד שולט הוא קטלוג 100% מהילדים המגיעים למערכת.

2.4.2 בשלב השני – חיזוי אילו ילדים היפכו למכירה בבדיקה של מעלה מ-70% (Accuracy).

2.5 עמידה במדדים

בחלק זה נתychס לכל ממד הצלחה בנפרד.

2.5.1 סעיף 2.3.1. הפיכת 30% מהילדים ל"ילדים חמימים" אשר יפכו למכירה

הפרויקט הצליח לעומת זאת כאשר 80% מהילדים ה"חתוכים" שסוגו באמצעות האלגוריתם הבלתי מונחה הפכו למכירה, וסה"כ 30% מכלל הילדים, (ולא רק מהילדים חמימים) הפכו למכירה.

2.5.2 סעיף 2.3.2. הגדלת רוחן הארגון ב-20%

מכיוון שאנו לא עובדים עם ארגון, וביססנו את הנתונים שלנו מתוך סקירת ספורות ועל Data Generator, אנחנו נוכחים שהצלחנו להגדיל את רוחן הארגון באמצעות השוואה לדוח רוח והפסד של חברת אלבר (קישור לדוח בראשית מקורות [16.4](#)).

לחברה אלבר בשנת 2021, היו 25,181 רכבים שהיו מיועדים להשכרה. סה"כ ההכנסות מהשכרת כלי רכב, שירותי מוסך, דרכר וగירה ומממן אשראי צרכני היה 1,293,751,000 שקל. עלויות שירות מוסך וניהול צי רכב וגירה הם כ: 000,201,000 שקל.

לכן סה"כ ההכנסות משירותי הליסינג הם 355,550,000 שקל. אם נרצה לחשב את ההכנסה מרכיב לחודש בודד, נחלק את מספר הרכבים מההכנסות, חלקו 12 חודשים ונקבל:

$$\frac{355,550,000}{12 \cdot 25181} = 1,176$$

סניף הליסינג שלנו מכיר רכבים בשווי של 7,166,044 שקל, אם נחלק את שוויו של כל רכב במספר הימים בהם הליקוי היה מעוניין להשכרה, נכפיל ב-30 ימים, ונחלק במספר הרכבים, נקבל שוויו המקורי שהשכרנו בכל חודש הוא 2,052 ש"ח.

כלומר הצלחנו להכפיל את השווי של השכרת רכב לחודש לעומת השווי של חברת אלבר אשר ממקמת מקום שלישי בחברות החזקות בשוק.

חסכנו זמן לאנשי המכירות על ניסיון מכירה לילדים חלשים ובכך יעלנו את הפעולות הארגון.

2.5.3 סעיף 2.3.3. מצורע איבוד ילדים פוטנציאליים בכ-80% באמצעות שימוש במערכת התראות לא יישומי ורכיב שינויים.

נציין שהפרויקט לא עמד במדד זה לא מחסור יכולת, אלא מכך שהחליטנו לשנות את אופיו הפרויקט ואת המטרה שלו.

2.5.4 סעיף 2.3.4. הסתגלות מהירה של משתמשי המערכת למערכת החדש וסיווג ילדים בצורה מהירה בכ-20% מהמערכת הקודמת

הפרויקט עמד במדד זה, המערכת שפותחה על בסיס אתר יכולת להיפתח בכל מכשיר בעל גישה לאינטרנט, בנוסף, קטלוג הילדים ושליחתם ללקוח לויקו סה"כ 25 שניות, גם לאלגוריתם הבלתי מונחה וגם לאלגוריתם המונחה.

לכן סיווג הילדים מתבצע בצורה מהירה ללא צורך בטמעה של המערכת אצל הליקוי.

2.5.5 סעיף 2.4.1. קטלוג 100% מהילדים לפי 4 קבוצות איות: ילדים רותחים, ילדים חמימים, ילדיםBINOMIIM, ולילדים קרימ

הפרויקט עמד במדד זה, המערכת מצילה לקטלוג את כל קבוצי הילדים המגיעים למערכת לארבעה קבוצות איות, ושליחתם ללקוח לאחר מיפוי הילדים.

2.5.6 סעיף 2.4.2. חייזי אילו ילדים הפכו למכירה בדיקות למעלה מ-70% (Accuracy).

הפרויקט עמד במדד זה, אחוז הדיקות שהגענו אליו עומד על 83%, פירוט על אחוז הדיקות ושאר הממדים של הלמידה המונחת בסעיף [11](#).

3 סקירת ספרות

3.1 מבוא לניהול לידים

3.1.1 ליד Lead

ליד הוא מושג בתחום ניהול המכירות אשר מייצג לקוחות פוטנציאלי המתעניין בשירות או מוצר המוצע למכירה על ידי עסק, חברה או ארגון. פרטיו ליד מכילים את פרטי יצירת הקשר עם הלוקו הפוטנציאלי המתקבלים על ידי אדם שפנה לאותו ארגון ישירות או אדם אשר השאיר פרטים ליצירת קשר בפלטפורמות השונות (דפי נחיתה, רשותות חברותיות ואתרי הארגון).

פרטיו המתעניין לרוב יכללו שם מלא, מס' טלפון וכתובת אימייל ליצירת קשר, פלטפורמה ממנה "נחת" הlid, גיל,מין וניתן להוסיף פרטים כמו מקום העבודה שבה עובד הlid ורכב בו התעניין להשכלה.

נהוג לפלח ולסמן לידים על בסיס מאפיינים דמוגרפיים וגיאוגרפיים וגם לסוג את מטרת הפניה של אותו הלוקו הפוטנציאלי כדי שיגיע לאיש המכירות המתאים והוא ייצור קשר עם אותו הלוקו הפוטנציאלי וינסה להניב מהשיכחה מכירה.

אנשים מכירות רואים את הלידים כמכירות פוטנציאליות להיות והлокו כבר הראה את התעניינותו בשירותי הארגון וכן חשוב מאוד לנוהל אותו וליצור קשר בחילוץ זמן קצר ככל האפשר. מטרת הפרויקט שלנו היא לעזור לאוותם אנשי מכירות לא Abed מכירות פוטנציאליות ולנהל אותם באופן חכם ויעיל על מנת להגדיל את אחוזי המכירות של אותו הארגון.

3.1.2 ציון לידים – Lead Scoring

ציון לידים הוא תהליך שנעשה כדי להבין את חשיבותו ועדייפותו של ליד עבור אותו הארגון. ציון הלידים עוזר להבין מה הסיכוי של אותה המכירה הפוטנציאלית להפוך להיות מכירה ממשית. החשיבות של הליד נקבעת על פי מספר פרמטרים שנקבעו מראש, לדוגמה: המוצר בו התעניין הלוקו, מצב דמוגרפי כמו תחום עבודתו (אם מדובר באדם פרטי), מגיל וכו'.

מודלי דירוג הלידים שמים דגש לדעתה מגוון כגון - מספר הקליקים אותו הליד הקליק על הקישור, מספר הפעמים שביקר באתרים השונים שלו בזמן ששוואה באתרים אחרים.

פעולה זו בתהליך של ניהול הלידים עוזרת לאוונן את האופי של הלוקוות העתידיים שלו ונעזר בהם בעת ביצוע המכירה.

במהלך ציון הלידים, ישנים אלגוריתמים רבים היודעים לזהות לידים שנראים מזוייפים ואנו הדירוג עוזר לאנשי המכירות לא לפנות לאותם הלידים ולהפסיק את זמנם.

במהלך הפרויקט שלנו נשתמש בשיטת ציון לידים ב כדי לדרג את המכירות הפוטנציאליות תוך שימוש אופטימלי בכל הדעתה אודות הליד.

3.1.3 משפר שיווקי – Sales Funnel

תהליך שעובר אדם מסוים מהרגע שהוא פוגש לראשונה את המוצר או העסק ועד הגיע שהוא מבצע את הרכישה הראשונה.

I. השלב הראשון: שלב החשיפה למוצר או לשירות של הארגון.

II. השלב השני שלב התעניינות והמחקר: לאחר שהאדם התרשם מהמוצר או מהשירות שנוחף אליו הוא יתחיל לברר עליו ויקים השוואת מחירים עם הארגונים המתחרים, יבדוק יתרונות וחסרונות וופולריות בשוק.

III. השלב השלישי שלב הבחירה: אחרי סיום הבדיקה והבחירה הלוקו ייצור קשר טלפוני או דרך המail או ימלא טופס באתר או בדף הנחיתה של המוצר עם פרטי ההתקשרות שלו.

IV. השלב הרביעי ריכשה: הליד הפרק להיות לקוח ללקוח רשמי של הארגון כאשר החליט לרכוש את המוצר או את שירותו הארגון.

V. השלב החמישי שימור לקוחות: שליחת עדכונים על מוצרים או מוצרים חדשים של הארגון שעשויים לעניין את הלוקו ולעודד להמשיך ורכוש את מוצר או שירות החברה.

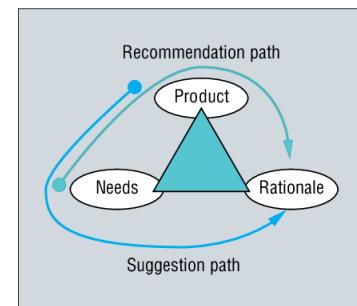
3.2 טכנולוגיה ויישום

3.2.1 מערכת המלצה

מערכות המלצה הן כלי תוכנה וטכניקות המספקות הצעות לפריטים שישמשו את המשתמש, ההצעות המוצעות נועדו לתמוך בהתהילך קבלת החלטות של המשתמש. פיתוח מערכות המלצה הוא מאיץ רב בתחום הכללי מתחומים רבים – כריית מידע, מערכות תומכות החלטה, בינה מלאכותית, סטטיסטיקה ועוד. המוטיבציה למערכות המלצה מוצגת בתרשימים 1, הרעיון המכני שמקורו בצריכים (בעיות) ולאחר מכן את הצורך, האלגוריתם החכם של הממליץ יכול להציג פתרון בפתרונות רבים שמבינים את הצורך, אומנם מערכות המלצה צרכות לאייר את הבעיה העיקרית של המשתמש, אך לא רק, מערכות המלצה ייעילות הן אלה שיאפשרו למשתמש לחזור את מרחב האופציות שלו ובהתקיים יתמכה בהעדפות השונות של המשתמש.

רוב משתמשי האינטרנט יכולים נתקלו אי פעם במערכת המלצה בדרך זו או אחרת, חנות ספרים וירטואלית שכאשר המשתמש יבחר ספר, תופיע לו הודעה "משתמשים שבחרו את הספר הזה, בחרו באותו הספר.." ניתן לראות, שכדי שהמערכת תוכל לתת המלצות עליה לדעת פרטימ אישים על המשתמשים שלה.

במסגרת הפרויקט תפוחה מערכת המלצה לאנשי מכירות מתחום הרכבים והלייניג על מנת לחסוך משאבם של זמן וcosa לארגון על ידיים עם סיכון הצלחה נמוכים, ולהתאים לאנשי המכירות לדיים בכדי שסיכון המכירה שלהם יעלן.



תרשים 3.2.1 הדגמה של מערכת המלצה

3.2.2 מערכת ניהול לקוחות – CRM

מערכת CRM הינה מערכת ניהול לקוחות הקשור ללקוחות המאגדת את מכלול כל ניהול הממוחשבים תחת מעטפת אחת ומציגת תמונה עדכנית ומלאה על הליקוח ופעולות הארגון בכל רגע נתון: כולל תהליכי מכירה, ניהול ושמור ללקוחות, ניהול פרויקטים, הנהלת חשבונות ומלאי, ניהול משימות באופן חכם, שליחת דיוור לרשימות תפוצה, ניתוח של נתונים ללקוחות באמצעות כלים של בינה עסקית וניהול תהליכי מכירה ומסעות שיוקן ופרסום.

מערכת CRM מדמה תמונה מלאה של הליקוח כולל מיזוג של כל פרטיו הקשר לתוך מסק אחיד ונגיש בקלות ונוועדה לשפר את איכות הקשר בין עסק או ארגון לבין הליקוחות שלו.

ארגוני, חברות ומוסדות רבים עושים שימוש במערכת כדי למן את הפוטנציאל העסקי מול הליקוחות, להגבר וליעיל את קצב העסקאות, לאגור מידע היסטורי להערכת הליקוחות, למן את היכולות השירותיות ולקבל תובנות עסקיות לטווח הארוך. לכל מערכת CRM יכולות להיות הגדרות ותכונות הייחודיות לה אך היתרונות המשמעותי שלה המערכת היא אחסון המידע אודות הליקוחות והאפשרות לשולף נתונים בצורה קלה ומהירה.

כאשר ארגון בוחר את מערכת ה-CRM המתאימה עבורו עליו להשוות בין המערכות השונות בשוק לפי התאמת לצרכים שלו - בין אם ניהול פרויקטים, מלאי, ביצוע מעקב זמינים או עבודת צוות או ניתוח מידע אודוט העובדים שלו.

בנוסף, במערכת ה-CRM נעשה גם שימוש לטובת ניהול לידים לעסקים והמערכת מסונכרנת עם הנתונים על הליקוחות הקיימים. בפרויקט שלנו אנחנו עתידים לחזור את השימוש במערכות ה-CRM השונות, השימוש בהן לטובת ניהול לידים והבדלים בין המערכות הקיימות בשוק לבין המערכת החכמה שאנו עתידים לפתח ולהבין את היתרונות והחסרונות של מערכת ניהול לידים אשר מתממשקת עם מערכת ה-CRM.

3.2.3 תכונות מונחה עצמים

בפרויקט נעשה שימוש בתכונות מונחה עצמים (או בקיצור OOP) לשם ייצור אובייקטים (עצמים) בעלי תכונות, מאפיינים ופעולות היכולים להתקיים כיחידה סגורה ועצמאית. תכונות מונחה עצמים נחשב לסת מוסכמת לכנתיבת תוכנה ומטאפיין כמודולרי. Möglich להפריד פונקציונליות של תוכנה למודולים עצמאיים הנינטנים להחלפה כך שכל אחד מהם מכיל רק היבט אחד של כל הפונקציונליות הנדרשת לביצוע. בנוסף, פרדיגמה זאת מאפשרת כמדרגית וניתן לתת לאובייקטים שהוגדרו יחס' היררכיות בניהם. תכונות מונחה עצמים מחקה את החשיבות בכך שהוא מסווים ומגדיר את העולם לקטגוריות אשר עוזרות לשЛОט ולסדר מידע ובו זמנית, תוך כדי הגדרת תכונות המיחדות כל קטgorיה.

בחרכנו להשתמש בתכונות מונחה עצמים כיוון שהיינו צריכים להגדיר אובייקטים שיתקשרו אחד עם השני אך בו זמנית לא יהיה ערבות בין טיפוסי עצמים שונים והאלגוריתם יהיה גמיש לשינויים בתוכנה.

3.3 למידת מכונה – Machine Learning

3.3.1 למידה בלתי מונחתית - Unsupervised Learning

למידה בלתי מונחתית (Unsupervised Learning) היא טכניקה בלמידה חישובית שבה מנסים ללמוד את התכונות והמבנה של אוסף דוגמאות נתונים כאשר הנתונים זמינים כפי שהם ללא תוספת תיוגים.

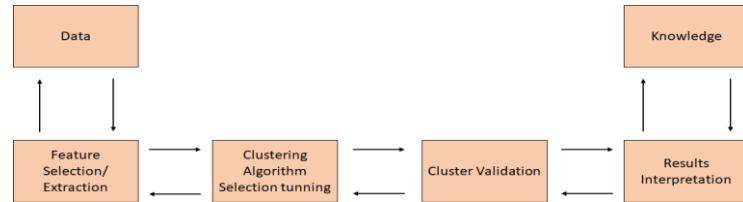
"ללא תוספת תיוגים (Labels)" מלומר שאנו לא יודעים את ה"תשובה האמיתית" ולאיזה קבוצה הדוגמאות הללו שייכות.

רוב הפעמים בעולם האשמי, כמו באתגר שהפרויקט שלנו עמד בו, המידע שלנו יגיע ללא תיוג לאיזה קבוצה הוא שייך, ונרצה לפתח אלגוריתם למידת מכונה שיאכל לסוג את המידע בצורה נכונה.

האלגוריתם עושה זאת בכך שהוא מוצא במידע ערכים חופפים ומחלק אותם לקבוצות, בנוסף הוא מזהה ערכים "אנומליים" שלא שייכים לשום קבוצה ומגדיר אותם ייחדי.

באופן כללי המטרה היא למצוא במידע תכונות מסוימות ולחבל אותם לאשכולות. בפרויקט שלנו, נרצה לחלק את המידע שלנו - לקבוצות פוטנציאליים (לידים) ולסוווג אותם לקבוצות.

התהליך שהמידע עובר בלמידת מכונה בלתי מונחת מוצגת בתרשימים הבא:



תרשים 3.1.1. התהליך שהמידע עובר בלמידה בלתי מונחתית

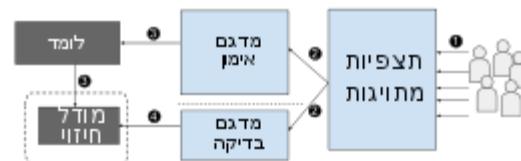
3.3.2 למידה מונחתית - Supervised Learning

למידה מונחית או למידה מפוקחת (Supervised learning) היא טכניקה בענף למידת מכונה, המאפשרת לפתח מכונה או מערכת (בדרך כלל תוכנית מחשב) שלומדת לפתור בעיות על בסיס מאגר גדול של דוגמאות "פתרונות".

אלגוריתם למידה אופטימלי הוא אלגוריתם שהפונקציה הנלמדת על ידי תוכל לחזות בצורה נcona את התשובה גם בעבר דוגמאות שטרם נראו על ידי המערכת.

המערכת בצורה שהיא נסורתה המשמש, האלגוריתם לומד תכונות מסוימות ו"מתאמן" על קבוצת האימון ולאחר מכן שנכenisו לו קלט תכמה שהוא לא יודע מה התיאוג שלה, לפי מה שהוא למד והוא יידע לאיזה קבוצה לשיך אותה.

בפרויקט שלנו, טכניקה צזו תבוא לידי ביתוי כבר בשלב הראשון בו אנו משתמשים בלמידה בלתי מונחית, אנשי המכירות ינסו לבצע מכירות על פי הקובץ המקורי. לאחר שאנשי המכירות ישתמשו בקובץ זהה הם ישלחו לנו את התוצאה הסופית והאם הצלחו לבצע מכירה או לא בכל Lid. המערכת תלמד מה מאפיין לקוח שהוא מעוניין ב מוצר, וכך תשפר את המלצות ההתחלהות אף יותר.



תרשים 3.3.2. דוגמה למערכת למידה מונחית

3.4 תהליכי preprocessing – עיבוד נתונים מקדים

3.4.1 Preprocessing – עיבוד נתונים מקדים

בפרויקטים בלמידת מכונה עיבוד מקדים של נתונים הוא שלב חשוב הקודם לתהליכי הניתוח של הנתונים, המאפשר ניקוי של הנתונים, בחירת מופעים, טרנספורמציה, בחירה של תכונות וכו'. שלב זה מאפשר גילוי של מידע לא רלוונטי ומיותר או נתונים רועשים העולאים לגרום לשיבושים בתפעול השוטף ולהחלטות שגויות בעיבוד הנתונים הסופי.

עיבוד נתונים מקדים כולל בתוכו 3 שימושים עיקריים:

1. ניקוי נתונים - בניקוי הדטה ניתן למחוק או למלא ערכים ריקים, מחיקת שגיאות נתונים לא רצויים.
2. אינטגרציה של נתונים - שילוב של נתונים מספר מקורות במקביל.
3. צמצום נתונים - לעיתים, בלמידה וטבלאות הנתונים מכילות הרבה מאוד מידע, מבצעים צמצום של הדטה וליקחת השדות והעמודות המתאימות ביותר תוך התחשבות בצריכי הפרויקט.

3.4.2 One-hot Encoding – שיטת One-hot

במעגלים דיגיטליים ובלמידה מונחית, שיטת One-hot-One מושתמש בשבייל לסמן קומבינציה של ערכים אשר ערך אחד מקבל את הערך 1, וכן השאר מקבלים את הערך 0 כחלק מתהליכי preprocessing.

בסטטיסטיקה ובלמידה מונחית משתמשים בשיטה זו בשבייל לייצג מידע קטגוריאלי בגלל שהרבה מודלים של למידת מכונה צריכים שהקלט של המידע יהיה מסווג נומירי, השתמש בשיטה זו בשבייל להמיר את המידע בתהליכי העיבוד המקדים.

אם למשל נקטalg את המידע הקטגוריאלי מ-1 עד N, יכול להיווצר מצב שבו צוין בדיוני שערך אחד מקבל באלגוריתם חשיבות גדולה יותר מאשר אחר.

לכן קידוד One-hot-One מוחל לעויתים קרובות על משתנים נומינליים על מנת לשפר את ביצועי האלגוריתם, ככלומר כל אחד מקבל עמודה משלה, וכל השאר אפסים.

אם מצב זה יכול להיות בעייתי, שם יש לנו יותר מדי ערכים שונים, יהיה לנו יותר מדי עמודות מה שיכל לעלות את רמת הסיבוכיות של האלגוריתם, לעלות את זמן החישוב ולפגוע בדיק, لكن נרצה להשתמש בשיטה זו כאשר יש לנו מספר סביר של ערכים קטגוריאליים.

3.5 אלגוריתמי המערכת

3.5.1 Logistic Regression

רגסיה לוגיסטי היא רגסיה בה המשתנה התלוּי הוא בינארי – כלומר, יש בו רק שתי קטגוריות (כמו "כן" ו-"לא", או "גבר" ו-"אישה"), המוצגות באמצעות הערכים 0 ו-1.

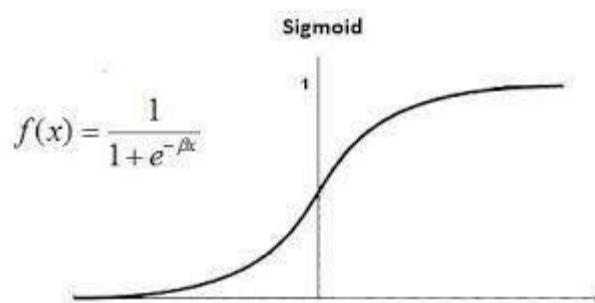
בשימוש במודל זה מתקבל פונקציה של ההסתברות לערך 1 כפונקציה של המשתנים הבלתי- תלויים, כלומר: מה הסיכוי שהערך של המשתנה התלוּי יהיה 1 כפונקציה של המשתנים הבלתי- תלויים.

נשתמש ברגסיה לוגיסטי כאשר המשתנה המוסבר הוא בסולם קטגוריאלי בעל שני ערכים בלבד. למשל, המודל משמש להבין כיצד מצב משפחתי וגובה המשכורת משפיעים על הבעלות על דירה. האם גובה משכורת או מצב משפחתי יכולים לנביא את הסיכוי שהנבדק בעל דירה?

הבעלות על דירה היא המשתנה ביןארי – במקרה זה, 0 מציין שהנבדק אינו בעל דירה ו-1 מציין שהנבדק בעל דירה, או במקרה שלנו 0 מציין שהליך לא רכש ליסינג ו-1 שהוא כן רכש ליסינג.

בסופו של דבר הפלט של הרגסיה על כל משתנה תהיה הסתברות בין 0 ל-1 שאפשר לתרגם אותה מה ההסתברות שהליך ירכש את הליסינג או לא, אם המספר יצא לנו הוא מעל 50% המודל ימיר אותו ל-1, ואם הוא מתחת הוא יחזה שהוא 0.

لتיאור מלא של אופן חישוב אלגוריתם רגסיה לוגיסטי בסעיף [9.3](#).



תרשים 3.5.1 דוגמה של הפונקציה הסיגмоידית שמשמשת את הרגסיה הלוגיסטיות

3.5.2 Decision Tree

עץ החלטה הוא מודל חיזוי בתחום הסטטיסטייה, קרית נתונים והלמידה החישובית המספק מיפוי בין תוצאות לערכים המתאימים עבורן.

עץ החלטה יכול לשמש כמודל חיזוי, המיפה תוצאות על פריט ויוצר מסקות על ערך היעד של הפריט. שימוש תיאורים יותר עבר עץ ההחלטה הם עצי סיווג או עצי רגסיה.

עץ ההחלטה הוא עץ ביןארי מלא המורכב מצומת החלטה שבכל אחד מהם נבדק תנאי מסוים על מאפיין מסוים של התוצאות, ועליהם המכילים את הערך החיזוי עבור התוצאה המתאימה למסלול שmobiel אליום בעץ.

סוגים של עצי ההחלטה הם עצי רגסיה שבהם מותאם ערך רציף לכל תוצאה ועצי סיווג שבהם מותאמים ערך בדייד או Class לכל תוצאה.

כמו כן קיימים עצי החלטה מסווג (Classification And Regression Tree) - CART המשלבים את שני סוגי החיזוי.

עż החלטה הוא ייצוג פשוט לSieog דוגמאות. למידה מbossה עż החלטה היא אחת הטכניקות הפופולריות והשימושיות ביותר ל-Supervised Learning באמצעות Sieog.

עż יכול "ללמד" על ידי פיצול קבוצת המקור לחתמי קבוצות, המתבססת על מנת ערך לתוכנה. תהליך זה של אינדוקציה מלמעלה למטה בעצם ההחלטה, הוא דוגמה לאלגוריתם נפוץ במיוחד לצרכים אשר לומדים לבצע החלטה מושכלת מהנתונים. בכרית נתונים, ניתן לתאר עż ההחלטה גם כשילוב של טכניקות מתמטיות וחויביות, המשמשות בתיאור, Sieog והכללה של קבוצה נתונה של נתונים.

لتיאור מלא של אופן חישוב אלגוריתם עż ההחלטה בסעיף [9.4](#).

Random Forest 3.5.3

יער אקראי, כפי שהוא מרמז עליו, מורכב מעץ ההחלטה. העצים נוצרים לרבות על ידי דגימה מהתמאנפיננסים או מהתמאנת הצפויות.

כל אחד מהעצים נותן תוצאה לא-אופטימלית אך באופן כללי, על פי רוב, החיזוי בדרך זו משתף. האלגוריתם של יער אקראיים בניית מקבץ של עצים רבים, כאשר בכל עץ בכל פיטול, הוא מגביל את המשתנים לפיהם הוא יכול לפחות- m משתנים בלבד.

(מתוך k אפשריים בדרך כלל $\sqrt{k} \approx m$)

כמו כן, האלגוריתם מגיריל צפויות (במקום להשתמש בכל הצפויות הוא משתמש במדגם שלו), לצורך בנייתו של עץ.

אלגוריתם זה יכול למצמצם את ההשפעות של קורלציה בין משתנים, וכך גם, הוא נותן הזדמנות למשתנים מסבירים שונים לבוא לידי ביטוי, אפילו אם הם לא בעלי העוצמה החזקה ביותר. לבסוף התוצר המתתקבל הוא ממוצע החיזויים על פני כל העצים.

עורך מוסיף לעצים הינו שnitן לחשב את ההפקטה הממוצעת במדד Gini של כל אחד מהמשתנים המסבירים, וזה מאפשר לדרג אותם לפי סדר חשיבות. בבעית גרטיה, החשיבות מסודרת לפי מידת "התהווות" (במנוחי RSS) שהוספה לשנתנה מסוימת תרמה לדיק, במשמעות.

لتיאור מלא של אופן חישוב אלגוריתם Random Forest בסעיף [9.5](#).

K – Means Clustering 3.5.4

אלגוריתם K-מרכזים (k-means) הוא שיטה פופולרית בתחום ה-[Unsupervised Learning](#) עבור ניתוח אשכולות (Clustering) בכרית נתונים.

מטרתו לחלק את הצפויות ל- k אשכולות לפי מרכז כובד (k-mean).

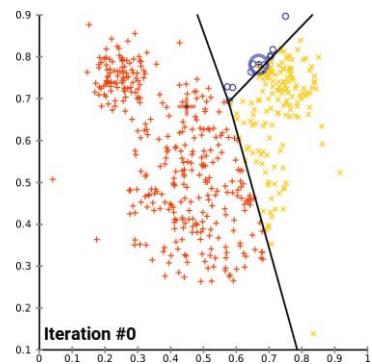
כל צפיפות משוכנעת לאחד מ"מרכז הכבד" ועל ידי בחירה נכונה של מרכז כובד ניתן לאתר את הקבוצות השונות.

נדרשות צפויות רבות על מנת להשתמש במודל ותוספת של צפויות חדשות עשויה לחיבר חישוב חוזר.

מדובר באלגוריתם היוריסטי שמשתמשים בו כדי לבצע חישובים שימושיים להתקנות לפתרון שאינו בהכרח הפתרון הטוב ביותר.

זהו מודל סטטיסטי שאינו מtabסס על ידע מוקדם על הנתונים אלא רק על צפויות בפועל.

لتיאור מלא של אופן חישוב אלגוריתם K - Means [9.1](#)



תרשים 3.5.4. מדמה איך אלגוריתם K-Means Clustering משנה את המרכזים בכל איטרציה וכך מחלק את הנקודות לקבוצות.

K - Modes Clustering 3.5.5

K-Modes הוא אלגוריתם למידת מכונה של במידה בלתי מונחת שימושים בו בשביל לחלק דאטה לקבוצות שבו אין לנו את התיאוג של אותו מידע, כלומר לאיזה קבוצה הוא שייר, בדיק כמו K - Means Clustering.

הבדל העיקרי בין K-Modes ל-K-means, שב-*K*-means משמשים במרחב אוקלידי בין כל איטרציה בשביל לחשב את המרחקים בין הנקודות למרכזים, ובכך מחלקים את הדאטה לקבוצות, ככל שהמרחב בין הנקודה למרczד שלה נמוך יותר כך יוכל להציג ביותר וודאות שהנקודה הזאת שייכת לקבוצה שאליה היא שייכת.

וב-*K*-Modes משתמשים בחישוב אי דמיון בין הנקודות, ככל שאי הדמיון נמוך יותר, כך הנקודות יותר דומות אחת לשניה, כלומר בעצם משתמשים במדד השכיח במקום מרחק אוקלידי. היתרון הגדול של שיטה זו שעכשו נוכל לחלק לקבוצות נקודות מסווג הדאטה הוא לא נומירי אלא הוא קטגוריאלי, כמו שיש בפרויקט שלנו, ובכך נוכל לחלק את הלידים לקבוצות גם אם הוא מכיל עמודות קטגוריאליות.

K – Prototypes Clustering 3.5.6

Prototypes Clustering – K הוא אלגוריתם למידת מכונה של במידה בלתי מונחת שימושים בו בשביל לחלק דאטה לקבוצות כאשר אין לנו את התיאוג של אותו מידע, כלומר, מה התוויות של אותו מידע, בדיק כמו K - Modes Clustering.

היתרון הגדול של אלגוריתם זה שהוא שילוב של שני האלגוריתמים האלה ביחד.

כאשר הדאטה סט שלנו מכיל גם עמודות שנן נומריות וגם עמודות קטגוריאליות ישנן שיטות שונות שנינן או להמיר את המידע הקטורייאלי לנומירי או להפוך, ואז ניתן להשתמש באחת השיטות עליה פירטנו לעלה.

אך הניסיון מראה, שבהרבה מקרים המטרה של העמודות פוגע באחוז הדיק של חילוק המידע, כי להפוך מידע קטורייאלי לנומירי ועל פיו לחשב את המרחק האוקלידי לא מסביר בצורה טובה את הנתונים, וגם להפוך.

ולכן לפי מאמרו של יאנג משנת 1997 (קישור ברשימה מקורות [16.4](#)), הוא הציע לשלב את 2 השיטות, וכשמדובר בעמודה קטגוריאלית חישוב המרחק בין המרכזים יעשה על ידי K – Modes וכאשר מדובר נומרית, חישוב המרכזים יעשה על ידי K – Means.

لتיאור מלא של אופן חישוב אלגוריתם Prototypes - K בסעיף [9.2](#).

3.6 מדדי דיק

Silhouette Score 3.6.1

ציון הסילווט היא מטריקה לביטוי רמת הדיק של "אשכולות של נתונים".

במיללים אחרים, זה מגדד הדיק שמשתמשים בו כאשר משתמשים באלגוריתם מסווג Clustering בلمידה בלתי מונחית בשביל להעיר עד כמה חלוקת המידע הייתה טובה.

הטכנית מספקת יציג גרפי תמציתי של עד כמה כל אובייקט סוג בצורה טובה, ככלומר עד כמה כל נקודה (במקרה שלנו כל ליד) דומה לשאר הנקודות שנמצאות באותו אשכול.

הציון האפשרי שיכול להתקבל הוא בטווח של [1,-1] כאמור:

- (+1) – החלוקה התבוצעה בצורה מושלמת.
- (-1) – החלוקה התבוצעה בצורה לא טובה.

לכן, נרצה שהמקדמים יהיו גדולים ככל האפשר וקרובים ל-1 כדי לקבל אשכולות טובים, אשכולות טובים אומרים שחלוקת הילדים שלנו לקבוצות התבוצעה בצורה טובה.

Confusion Matrix 3.6.2

טבלת Confusion matrix משמשת לביעות של סיווג מונחה על מנת להראות בצורה ויזואלית האם המודל שלנו חזה בצורה נכון. ואפשר על ידה להבין ממנה כמה מדייקים בכל סיווג, וכמה טעו בחיזוי.

		PREDICTED	
		POSITIVE	NEGATIVE
ACTUAL	POSITIVES	TRUE POSITIVES	FALSE NEGATIVES
	NEGATIVE	FALSE POSITIVES	TRUE NEGATIVES

איור Confusion Matrix 3.6.2

Accuracy Score 3.6.3

"דיק" הוא היחס בין הערכים החיוביים והשליליים שנחצז נסך כל הערכים הק"י'מים במאגר המידע. דיק גבוה מעיד על טיב המודל, ככלומר כמה הוא הצליח בסה"כ לחזות נכון את הערכים החיוביים והשליליים.

בפרויקט שלנו בחלק השני של האלגוריתם הלמידה במונחית, אנחנו נעירר האם אנחנו מצלחים לחזות אילו ילדים הפכו למכירה באמצעות אלגוריתמים של למידת מכונה.

נעירר האם האלגוריתמים שלנו חזו טוב באמצעות Accuracy Score שمراה לנו כמה מדויק היה המודל.

$$Accuracy = \frac{TP + TN}{N}$$

Precision Score 3.6.4

מדד ה-"דיקנות" הוא ממד אשר עונה על השאלה הבאה: "איזה חלק מהנקודות החשובות החזוויות הוא חיובי באמת?"

מדד הדיקנות מחושב על ידי סכמתה של כל הילדים שנחזו שהם הפכו למכירה ובאמת נמכרו (TP), חלקו כל מה שנחזו שהוא חיובי (כלומר כל אלה שבאמת נמכרו) ועוד אלה שנחזו שהם נמכרו והם לא נמכרו (FP).

המודל ביצע חיזוי איזה ליד יփוך למכירה בחלוקת הוא החלט לחזות ובחלקם הוא טעה. לכן נרצה לדעת מכל הילדים שנחינו שהם הפכו למכירה, כמה מהם באמת הפכו למכירה. מדד זה טוב כאשר אנחנו רוצחים שהמודל שלנו יהיה כמה שיותר מדויק, בשביל לא לשבץ זמן (וגם בסוף) לאנשי המכירות עם לידים שכביכול יש להם פוטנציאל מכירה גבוהה, אך לבסוף הם לא הומרו למכירה.

TRUE POSITIVES

TRUE POSITIVES + FALSE POSITIVES

Recall 3.6.5

Recall מדמה את התשובה אודות: "איזה חלק מההתוצאות בפועל זהה בצורה נכון?" מדד ה-Recall מחושב על ידי סכמתה של כל הילדים שנחזו שהם הפכו למכירה ובאמת נמכרו (TP), חלקו, כל הילדים שנחזו שנמכרו ובאמת נמכרו (TP) ועוד הילדים שהיו אמורים להיות חשובים שנמכרו, אבל הם נחזו שהם לא נמכרו (FN).

מדד ה-Recall הוא ממד חשוב שנותן אינדיקציה טוביה האם המודל באמת חזה נכון איזה ליד הפרט למכירה.

לדוגמה: אם בדата סט האימון רוב התוצאות בעמודת המטריה הם 0, ויש רק תוצאה אחת שהיא 1, אז המודל שלו יתאמן על הדטה סט הזה, ומכיון שרוב התוצאות היו 0, ברגע שנכנים לו קלט חדש, הוא יחזה בסיכוי גבוהה שהיה 0. דבר שנקרא דата לא מאוזן (כלומר שאין יחס שווה בין תוצאות החשובות לשיליות) ולכן Recall יהיה מאוד נמוך.

Recall גם יכול להראות שהדטה לא מספיק מאוזן, וגם יכול להראות שלמרות שקיבלנו גובה, זה לא בהכרח אומר שהמודל חזה בצורה נכון. Accuracy

TRUE POSITIVES

TRUE POSITIVES + FALSE NEGATIVES

F1 – Score 3.6.6

מדד F1 הוא ממוצע משוקל של ה-Precision וה-Recall, ובכך מחשב ומשקל את השגיאות משלני הסוגים הללו.

הчисוב הוא בהתאם לנוסחת הממוצע הרמוני.

הנוסחה לחישוב ה-F1-Score:

$$F1 score = \frac{2 * Recall * Precision}{Recall + Precision}$$

ממוצע הרמוני פירשו תחילה לממוצע פשוט, מכיוון שהוא כולל נתונים קיצוניים כלומר אם עמודת המטריה לא מאוזנת.

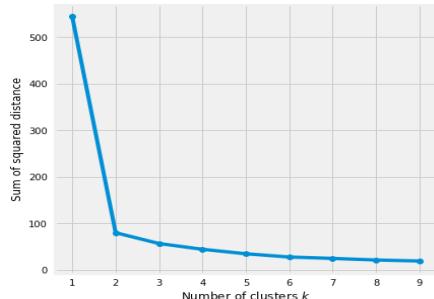
למודל שקיבל תוצאה עם Precision של 1 ו-Recall של 0 יש ממוצע פשוט של 0.5 אך הציון-F1 Score שלו 0.

ציון F1 נותן משקל שווה לשני הממדדים. אם נרצה ליצור מודל סיווגamazon עם איזון אופטימלי של F1-Score, אז ננסה למקסם את ציון Precision-Recall.

3.6.7 Elbow Method – שיטת המרפק

שיטת המרפק נותנת לנו מושג מה יהיה מספר k טוב של אשכולות בהתבסס על סכום המרפק הריבוע (SSE) בין נקודות הנתונים והמרכזים של האשכולות המוקצים להם. נבחר את k במקומות שבו SSE מתחילה להשתטח וליצור צורה של מרפק.

לדוגמה:



תרשים 3.6.7 תרשימים שיטת המרפק

ניתן לראות באופן ברור ש- $\text{SSE} = K$ ה- K -E-SSE יורד ממשמעותית, ומ-2 הוא נשאר בערך באותה רמה,

לכן ניתן להסיק ש- $2 = K$ בדוגמה זו היא בחירה הגיונית ומתאימה.

בפרויקט שלנו השתמש בשיטה זו בשבייל להחלטת כמה קבוצות של ידיים לחלק, כיצד נתיבים קבוצה זו (ידיים חמימים, ידיים בינוניים, ידיים חלשים, וכו').

3.7 בסיסי נתונים - Databases

בשנים האחרונות, לאור עלייה השימוש במחשב בענן ומtran אפשריות לאחסון נתונים בנפחים גדולים מבלי להחזיק שרתים פיסיים נוספים, המונח Big Data מקבל משמעות וצבר תאוצה.

החלומות מכריינות דורשות מידע אמין שמקורו בעבודות גולמיות בכך לוחבל לשינוי מגמות חיוביות. אחסון נפח נתונים בנפחים גדולים, מאפשר לבצע ניתוחים סטטיסטיים המסתמכים על מידע רב ובכך עשוי לשפר מודיע דיווק בעת ניוח הנתונים והפקת התובנות מהם. בסיס הנתונים הוא אמצעי המשמש לאחסון מסודר של נתונים בשורת, נתון המקל על פעולות שליפת נתונים, ניהול וערכונם.

באמציאות בסיס הנתונים, יוכל להסיק מידע רלוונטי ומדויק לגבי הנתונים הגלומיים העומדים לרשותנו, להפיק תובנות ולסייע בקבלת החלטות מכריינות בארגון. ניתן להשתמש ב-DBMS (Data Base Management System) לבניית סוגים רבים ושונים של מאגרי מידע בהתאם לאופי הארגון.

כיום, ישנו שני סוגים מרכזיים של בסיס נתונים:

3.7.1 בסיס נתונים רלציוני – Rational Database Management System

מודל בסיס הנתונים הוויטק מבין השניים. במודל זה, הנתונים מנוהלים בטבלאות, כשלכל רשומה מתוארת על ידי שורה וכל שדה מאופיין על ידי עמודה. לכל טבלה יש שדה מזהה ייחודי המאפשר גישה וממשק לטבלאות נוספות בסיס הנתונים. הנתונים מאופיינים בסוג (Type) מוגדר מראש. מודל זה עלול לצרוך נפח אחסון גדולים יותר בשל מוגבלותם לבניה הטבלאות ומורכבותם בשימוש נתונים אשר במבנה שאינו רלציוני (XML, JSON).

3.7.2 בסיס נתונים שאינו רלציוני Non Rational Database Management System

בסיס נתונים מסווג זה נקראים גם NoSQL (not only SQL) Database. מודל זה מאפשר לאחסן נתונים במבנה מורכב יותר, אך עם זאת יעיל וחסכוני בהתייחס לנפח אחסון.

מודל בסיס נתונים שאינו רלציוני מאפשר לאחסן אוצרות נתונים תחת נתון ספציפי, ולאפשר חלוקת הנתונים בסיס הנתונים כך שזמן התגובה יצטמצמו.

מודל NoSQL מאפשר ניהול מסדי נתונים בנפח גודלים מאוד ועיבוד מיידי רב במקביל על פני מערכות מבוזרות רבות. מלבד מחשוב ענן, מסדי נתונים של NoSQL מתמחים ליישומי Web 2.0 וברשתות חברתיות, כאשר קנה המידה האופקי כולל אלפי צמתים. מקורה היא במערכות פיתוח של חברות Google ו- Amazon.

SQL Lite 3.7.3

מסד נתונים ייחסי (רלציוני) המפורסם קוד פתוח ווՈולרי בזכות היותו תהליך עצמאי נפרד ואין תלוי בשרת המופעל מטה ליר קיים ומשמש בעיקר כדי לקוח שירותי ניהול מקומי.

קובץ הרצה שלו שוקל מעט אך תומך גם בנפחים גדולים של מידע וכל מסד נתונים בו מנוהל בתוך קובץ אחד במערכת הקבצים. מסד נתונים זה בעיקר משמש אפליקציות בזכות כל היתרונות הטמוןים בו.



3.8 שירות אחסון

3.8.1 שרת פייזי

ספק שליטה מוחלטת על המידע ואני רגש לתקלות בחיבור לרשות, לצורך גישה לנתונים קיימים, אך דורש תחזוקה רציפה. שרת פייזי מוגבל בפריסתו ולכן חשוף לאיבוד מידע רב יותר בעת תקללה. תהליך גיבוי המידע מורכב וארוך.

3.8.2 מחשוב ענן

ספק שליטה על הביצועים ונפח האחסון בהתאם לדרישה, אך העלויות עליה בהתאם. שרת ענן יאבך מינימום מידע בעת תקללה נכון העובדה שהגיבוי זמין, מהיר,iesel ויכול להתבצע במרוחchi זמן קצר. המידע נגיש מקורות שונים ואני מוגבל לתחנת עבודה ספציפית, אך כאשר תיהה תקללה באינטרנט, המידע אינו יהיה זמין.

3.9 שפות תכנותצד לקוח:

3.9.1 שפת תכנות HTML:

שפה תכנות המשמשת לבניית עמודי תוכן אינטרנט הכול含 הגדרת פסקאות, כוורות והטמעת תמונות ורטטונים בדפים. ביום זהה השפה העיקרית בה משתמשים היום והוא פשוטה ללימוד, ניתן לפתחה במהירות, ומטאפיינית ביכולות נגישות גבוהה, כשניתן להתממשק עם שפה זו מכל מחשב.



CSS-Cascading Style Sheets 3.9.2

באמצעות שפת תכנות HTML למדנו כיצד מתכוונים מכניםים תוכן בסיסי לרשות. עם זאת, חיבים להוסיף צבע, חיות וסוגנון לדפי אינטרנט ובפורמט נפוץ המופיע בכל הדפסנים והמכשרים הדיגיטליים. זה המקומם בו CSS נכנס לפעולה וمبיא לחיבים את האסתטיקה ברשות.

כל התוכן המורכב שנוצר ב- HTML הופך לאמנות חזותית וקובע את "הזהות" של דף האינטרנט. מטרת CSS היא להפוך את האתר לצבעוני עם פריסה יצירתיות וסוגנונות פונטיים.



3.10 שפות תכנות שכוללות כלים למטרת קלסיפיקציה

Python 3.10.1

שפת קוד חינמית הידועה כקללה לשימוש, בזכות התchapיר הנוח שלה. בנוסף, פיתון ידועה שהיא שפה שנוכה לשימוש בעולם Data Science מכיוון שיש בה המונ ספריות מובנות שהופכות את תהליך קרית המדע לפחות ומהר.

ספריות עיקריות שבהם נעשה שימוש: scikit-learn, Pandas, NumPy, Matplotlib, Seaborn (פירוט על הדוקומנטציה של הספריות העיקריות בהן עשינו שימוש בסעיף [16](#)).



Jupyter Notebook 3.10.2

מסמכים מחברת הם מסמכים המיוצרים על ידי אפליקציית Jupyter Notebook, המכילים קוד מחשב (למשל פיתון) וגם רכיבי טקסט עשיר (פיסקה, משוואות, דמויות, קישורים וכו').

מסמכים מחברת הם מסמכים הניתנים לקריאה על ידי אדם המכילים את תיאור הנתונים ואת התוצאות (איורים, טבלאות וכו') וגם מסמכים הפעלה שניתן להפעיל לביצוע ניתוח נתונים.

לפני הקוד המתעסק בלמידת מכונה, ביצענו ניתוח סטטיסטי ויזואלייזציה על הנתונים בשביל להחלטת באיזה מודל להשתמש, מה הינו התוצאות שלו, ועוד מידע נוסף אשר גורם לנו להבין Aiזה ליאז הפך למכירה.

הוספנו קישורים למחברות Jupyter נעל שכתבנו בנספחים, והוספנו אותם גם כקובצי PDF בצורה נוחה לקריאה שהיא אפשר לעבור על המחברות השונות שכתבנו בסעיף [17.6](#).



PyCharm 3.10.3

PyCharm היא סביבת פיתוח משלבת ייעודית של Python המספקת מגוון רחב של כלים חיוניים למפתחי Python, המשולבים באופן הדוק ליצירת סביבה נוחה לפיתוח פרודוקטיבי לPYTHON, אינטרנט ומדעי נתונים.

בשביל להcin את הסקריפטים השונים ואת האטר עצמו השימושו בסביבת הפיתוח זו אשר נוחה לשימוש, חינמית, ושהתנסינו בה במהלך הלימודים בתואר.



4 מצב קיימן

4.1 סקירה ותיאור מצב קיימן

כiom בסוגיות ניהול הליסינג חברות הליסינג השונות בארץ ובעולם רוב החברות עובדות בצורה מושנת כאשר הלקוח מעוניין לדבר עם נציג מכירות.

הילדים עוברים בצורה אוטומטית לאחראי המכירות דרך אפליקציית WhatsApp מהפלטפורמות השונות (פייסבוק, אחר החברה, השארת פרטיים בצורה טלפונית, רשות חברות ווכ'ו), והוא שולח כל ליד לפיה התchrom והעיסוק העיקרי לאיש המכירות המתאים ומוסיף את הליד בצורה ידנית לקובץ Excel מאוחד. על בסיס סקר הספרות שביצענו ראיינו כי דבר זה גורם לפספוס לדיים רבים ותיכנן לא עיל של זמני העבודה.

ישנם פתרונות נוספים המציעים ניהול לידיים חכם ויעיל אך פתרונות אלו דורשים הטמעה מלאה של המערכת במערכות הארגוניות של החברה (כמו Salesforce וכו').

4.1.1 חברות הליסינג בישראל

באיזה, מחזיקות חמישה חברות הליסינג הגדלות במעלה מ-200 אלף מכוניות, כאשר "קבוצת שלמה", "Hertz" ו"ו.ט.א.ס" הן שלוש חברות הגדלות והשלכות בתחום חברות הליסינג והשכרת הרכב.

"שלמה סיקסט" רשמה בשנת 2021 הכנסות של מעל 3.8 מיליארד שקל והעסקה כ-1,393 עובדים. בשנה זו החברה החזיקה בכ-77 אלף כלי רכב.

אחריה בדירוג נמצאת אלבר, שרשמה הכנסות של 2.8 מיליארד שקל, העסקה 1535 עובדים והחזיקה בכ-50 אלף כלי רכב.

ו.ט.א.ס, המדורגת במקום השלישי, הכנסה כ-1.5 מיליארד שקל, העסקה כ-758 עובדים והחזיקה ב-2020 כ-36 אלף כלי רכב.

חברות נוספות נוספות הן אוויס, אלדן, באדגט, ליסקאר, Lease4u ו-enterprise.

שם החברה	דוחות 2021	מזהר 2020 (ב מיליון ש"ח)	מזהר 2020 (ב מיליון ש"ח)	מזהר 2020 (ב אלפי ש"ח)	מבנה עיקרי	שם יסוד	שנת יסוד	דוחות בענין 2020				
קבוצת שלמה	1					3,879.0	68	10	עתליה שמלצר	1	1974	1
אלבר שירותי מימוניות	2					2,840.8	95	17	קבוצת אלעדרא	2	1994	2
ו.ט.א.ס - יוניברסל פתרונות תחבורה▲ (Avis)	3					1,597.2	153	25	UMI	4	1986	4
קשר רט א קאר HERTZ ישראל	4					1,505.1	167	30	יעקב ונילי שחף, ישראל קז	3	1985	5
אלדן תחבורה	5					1,410.4	174	31	משפחודהן	5	1967	5

טבלה 4.1.1 חממת חברות המובילות בשירותי הליסינג בישראל.

היום חברות הליסינג בישראל עובדות בכמה מתכונות שונות כאשר מדובר בהשכרת רכב להשכרה, יש הבדל האם הלקוח מעוניין בהשכרת רכב לטוויה הקצר, או השכרת רכב לטוויה הארוך (לייסינג) נראה כיצד המשך נראה בכל אחת מהמתכונות באמצעות דוגמה מהחת חברות הליסינג באתר האינטרנט שלה.

4.1.1.1 השכרת רכב לטוויה הקצר:

ניתן להשכר רכב דרך האתר החברה לטוויה הקצר כאשר המשתמש בוחר בדיקן באיזה רכב הוא מעוניין, מפרט את תקופת ההשכרה, מהייןIASOF את הרכב ויעד החזרתו, לאחר הכנסת נתונים אלו, המערכת באופן אוטומטי שולחת הצעת מחיר למשתמש.

דוגמה דרך אתר שלמה SIXT:

השכרת רכב בישראל

הקלideo יעד לאיסוף והחזרה

הציג מחרים

המשתמש בוחר את הארץ
איסוף והחזרה

נמל תעופה רמון אילית
נמל התעופה ע"ש אילן רמון, 18 ק"מ מאיילת

17:00 - 08:00 א' - ה'
12:00 - 08:00 ו' - ש'

נמל תעופה רמון אילית

סכינים בעיר

איילת ■
 אשדוד ■
 באר-שבע ■
 בית שמש ■
 גן גבורה ■

שלמה זקס

לוואין ■
 נפה ■

הசזרה נספחת בטלפון: 050 910 0000 | סמסונג גלקסי S21 | מושג עד 10%

Map data ©2021 Google | מושג עד 10%

Google

איור 4.1.1 תצלום מאתר "שלמה SIX" – דוגמה להשכרת רכב לטוווח הקצר מחברת Ark-Chipズ'ו שפונקציונליות זו אפשרית כאשר מדובר בהשכרת רכב לטוווח הקצר, כאשר מדובר בהשכרת רכב לטוווח הארוך (לייטיניג) הפונקציונליות והתהיליך משתנה בהתאם.

4.1.1.2 השכרת רכב לטוווח האחרון:

כasher madbar bahechraat rachb klyisng (hescraa letwoh haaror) hopenkzionlit baatir meshana

מגון דגמי 2022 בליסינג פרטיפיעול!



מכירת רכב מיד רашונה

ליסינג פוטו - חדש פרטיפיעול!

0 ק"מ רכב חדש

חפשו רכב

בחר דגם

שומם

טען-טענה

מלווי לט טקטי חודשי: חטף הכל!



טווינקה יאריס סול 1.5 אוטו!
מחיר נקיון | 2022 | 0%

₪ 1,899⁰⁰

מחיר כולל מסים
מחיר כולל מסים
מחיר כולל מסים

כאשר חיפשנו רכב ליסינג, המערכת מציעה לנו לבחור את סוג הרכב בו אנחנו מעוניינים (רכב משפחתי, ג'יפ, מסחרי וכו'), יצרן, ודגם.

לאחר שבחרנו את מאפייניהם אלו, לדוגמה "טויוטה יאריס" אתר החברה מעביר אותנו להשارة פרטים באתר.

טוויטה יארס סול' 1.5 אוטו', 2022

פרטי פגעה של שלמה	侃יה	ליסינג פרטי תעשייתי
מבחן 1,899 לחדש	שנ' 105,900 יבואן	מבחן
36 חדשים	חומרן	תנאי תשלום
שנ' 8,900	-	תשלום ראשוני
✓	-	פסוי בטוח/חייבת אחריו
✓	-	טיפולים
כ-3 חודשים	آن צורך למכוור את הררכ	מכרות הרוכב
עד 45% בשלוש שנים	ירידת הערך עליינו	ירידת ערך

בכפוף לתמליא ומועד חיבור הבואן

מעוניינים בבדיקה? עכשווי נס בפורטת וידאו עם נציג

המשר
לlezat טרייד אין
לרכב הישן שלך
אינ'
צוז אידי' קשור
סניף
טלפון
שם

וכאן נכנס לתמונה נציג השירות של הסניף והפתרון שאנו חenso רוצחים להציג בפרויקט, לא ניתן לבצע חזהה ליסינג דרך אתר האינטרנט ואנו זוקרים לנציג שירות על מנת לבצע את ההשכרה. הפרטים שלנו מעוברים בצורה אוטומטית לקובץ אקסל שנציג המכירות אמרור לחזור לידים האלו בשיטת FIFO ללא מילון וסיווג של הלידים.

4.1.2 אופי התהיליכים הקיימים

לאחר שביצענו סקירה על הדרכים השונות בהן ניתן לשכור רכב דרך אתר החברה ולמרות גודל השוק של חברות הליסינג בישראל (סך ההכנסות השנתיות מפעילות ליסינג לשנת 2021 מוערך בכ- 12 מיליארד ש"ח), הגענו למסקנה פשוטה, עדין רוב חברות הליסינג היום עובדות בצורה "לא חכמה".

"לא חכמה" הכוונה לכך שתפקיד הלקוחות הפוטנציאליים שהם משתמשים דרך שיווק דיגיטלי (רשותות חברותיות, פרסום/google וכו') או דרך אתר החברה, הנציגים פונים ללקוחות הללו בשיטת "פיפו" (First In First Out) ואין סינון או תיעודף של לקוחות פוטנציאליים, لكن ניתן להגדיל בביטחון שהטהיליכים בחברות אלו רק "נותנים מענה"/שימוש ב-AI ו-*Lead Scoring* הוא דבר מתבקש על מנת להגדיל את רוחן הארגון לחסוך זמן ומשאבים לעובדים.

4.1.3 הנהלים וסטנדרטים קיימים

מנהל סניף מגדר את הנהלים לכל סניף. לאחר בדיקה עמוקה של הנושא במספר סניפים שונים, רأינו כי כל הסניפים מקפידים על ניהול ידע ועובדים על פי הנהלים מסוימים שמוברים לכל עובד חדש בחברה הדורשים תיעוד מקיף.

כאשר נציג מכירות יוצר קשר עם לקוח פוטנציאלי, הוא דואג לתעד את תשובה הלקוח בקובץ Excel עם כל פרטי הליד, ומוסיף גם פרטים רלוונטיים בהתאם לשיחה. במידה ומתרבעת מכירה – העובד פותח ללקוח תיק ללקוח במערכת של החברה ובها הוא רושם את כל הפרטים הנספחים של המכירה.

4.1.4 הגורמים הפנימיים והחיצוניים בעלי השפעה על התנהלות הארגון

טבלה הבאה מתארת סניף של חברת ליסינג:

גורם	יחסו הgomlin עם הארגון
פנימיים	מנהל סניף לכל חברות הליסינג ישנים כמו סניפים הפרושים ברחבי הארץ, מנהל הסניף אחראי על ניהול העובדים, מעקב אחר המכירות, לעומת זאת הסניף עוד
	ארגוני מכירות אנשי המכירות של הסניף, תפקידם להיפגש עם לקוחות מסוימים המגיעים לסניף, לבצע מכירות טלפוניות באמצעות לדיים המגיעים דרך שיווק דיגיטלי, להתאים את הרכב הנדרש ללקוח.
	מנהל מכירות לכל סניף יש מנהלי מכירות שבאחריותם לקבוע יעד מכירות, לנוהל את אנשי המכירות ואת המכירות באופן שוטף, מחלק את קובץ הלידים לאנשי המכירות.
	איש אדמיניסטרציה קובע פגישות לאנשי המכירות, מדפסה ומתייגת חזים שנחתמים בסניף, קובע את ימי החזרת הרכבים ולקיחתם, אוסף את הלידים מפלטפורמות שונות מעבירותיהם למנהל המכירות.
	נהג שינוי שינוי רכבי הסניף – לקיחתם לשטיפה, תדלוקים, טיפולים, העברת טסטים, סידור ושינוי הרכבים במגרש הסניף, מתן שירות לקוחות.
	מחלקה שיווק אחריות על קידום מכירות ומיוגר הארגון. פרסום הארגון בשרותות חברותיות, ניהול האתר של חברת הליסינג, לארגן ולסדר את הלידים וליציאתם לאנשי המכירות.
חיצוניים	צרכים פרטיים רכישת רכב דרך חברת ליסינג, כולל תשלום של מקדמה ותשלומים קבועים מדי חודש. בהתאם לתקופת הרווח יכול לבחור בין תשלום יתרת המחיר וההעברה בעלות על הרכב לרוכש, מכירה של הרכב חוזרת לחברת הליסינג ו/או רכישת רכב חדש במסלול דומה.
	צרכים עסקיים מעסיקים שורצים לחסוך במשאבים בניהול צי הרכב ולבצע מיקור חזע עבור צי הרכב שלהם, וגם עבור מי שורצה לתת לעובדים הטבה משתלמת מאוד.
	ספקים הספקים לחברות הליסינג, אחראים לספק את אלפי הרכבים לחברות, מדובר על חברות גדולות הנמצאות בחו"ל.

טבלה 4.1.4. תיאור סניף ליסינג

4.2 סקירת המצב הכספי בארץ ובעולם

4.2.1 סקירת המצב בארץ

העובדים מנהלים את הלידים בקובץ Excel, משמע אין סync' בין העבודה של העובדים- יכול להיגרם מצב שבו שני נציגים מסוים מתקשרים ללקוח במקביל- דבר שיכול להשפיע על החלטת הלוקו שמקבל הרגשה של חוסר מקצועיות מעובדי החברה.

בנוסף, העובדים מקבלים את קובץ האקסל לא מסודר לפי איזהם פרמטרים שיכולים לעזור לדרג את רמת חשיבות הלוקו, ואת חשיבות החזירה המיידית אליו.

סקר השוק שעשינו בארץ כלל את חברות הליסינג המובילות בארץ אשר פועלות באותו מתחום גם בחו"ל. מבדיקה עם אנשי המכירות ועם מנהלי המכירות נאמר כי במדינות שבהן חברות עובדות התהילכים מtbody"ם באוטה הדרך.

ההבדל הגדול ביותר בהתנהלות השוטפת של חברות בארץ ובחו"ל הוא שלוקו עסק' בחו"ל בעט مليי' פרטיא הליד, הלוקו הפוטנציאלי מתבקש למלא את הקילומטראץ' שהוא צופה שיעשה במהלך השנה - דבר שעזר לקביעת מחיר הליד.

4.2.2 סקירת המצב בעולם

סקירת המצב הכספי בעולם תואם למצב הכספי בארץ.

אר, בשונה מחברות בארץ, בחו"ל ובדגש על אריה"ב לפניה ביצוע עסקה בתחום הליסינג, חברות בודקות את דירוג האשראי של אותו הלוקו הפוטנציאלי.

בכדי לקבל את האישור באלה"ב הדירוג הוא 680 ומעלה, וחברות הליסינג שמות דגש על הללואות רכבים היסטוריות בכך להבין אם אותו הלוקו יוכל להחזיר בזמן את תשלומי הליסינג עבור הרכב.

במידה והדירוג הוא מתחת ל-680, הלוקו כן יוכל להסביר את הרכב אך יctrkr לשלם עבור השירות הרבה יותר.

בכדי לבצע את בדיקות האשראי, על החברה ליצור קשר עם אחת מ-3 חברות דיווח האשראי, ולרבות חברות האשראי גבות בין \$ 25-\$ 75 עבור בדיקה ללקוח פוטנציאלי בחברות קטנות, עבור חברות גדולות ישנים הסכמים עם חברות דיווח האשראי המספקות הנחות עבור בדיקות אלו.

היות ובדיקות האשראי הן הכרחיות הן מבחינה פיננסית והן מבחינה חוקית, חברות הליסינג בחו"ל מספקות רכבים רק למי שיש לו דירוג מתאים, או מעילות את מחיר הליסינג בהתאם. لكن, בעט דירוג ליסינג בחברות באלה"ב אחד הפרמטרים החשובים שנשים עליו דגש היה דירוג האשראי ועמו זה בקובץ הלידים קיבל משקל גבוה יותר בעט שקלול המשתנים.

5 מסמך דרישות

מסמך הדרישות ושינויים מפורט בסעיף [15](#).

6 בחינה וניתוח חלופות מערכתיות

6.1 אלטרנטיבות לשימוש הפרויקט

6.1.1 חלופה 1 - מערכת ייעודית בפיתוח עצמי מלא

פיתוח מלא של מערכת המלצה ייחודית עבור חברות הליסינג וכחיתבת אלגוריתם המבוסס על הדרישות הפונקציונליות של הפרויקט. לטובות בנית המערכת יעשה שימוש בשפת Python שפת תכנות דינמית שתאפשר לנתח את הנתונים אשר יאחסנו בשרת ייעוד. ניתוח המידע וביצוע האנליזות יבוצעו גם על ידי שפת Python בסביבות פיתוח התומכות בשפה זו.

יתרונות חלופה 1:

- חוסר תלות בגורם חוץ לצורכי פעולה ושיפור המערכת.
- מידת התאמת של המערכת לצרכים גבוהה מאוד.
- שימוש בסביבת עבודה מוכרת, חינמית, בעלת שם עולמי ובעל ספריות מובנות לביצוע ניתוח הנתונים.
- גמישות גבוהה לשינויים תוך כדי פיתוח ושיילוב שינויים עתידיים במערכת.

חסרונות חלופה 1:

- הקמת מערכת ופיתוח שליה תדרוש זמן רב.
- פעולה שוטף ותחזקה עצמאית של המערכת.
- יתכנו עלויות נוספות עקב ייעוד מאנשי תוכנה מתקדמים לטובת מתן אבטחת המערכת.

6.1.2 חלופה 2 - רכישת תוכנת מדף וביצוע התאמות לצרכי הלקוח

חלופה זו כוללת פיתוח של מערכת לשילוב של מערכות קיימות. מערכת Choin Einstein הינה توוסף למערכת ה-CRM של חברת Salesforce והוא ייחד מספקות פלטפורמה לביהול קשרי לקוחות לארכונים. מערכת Einstein היא מערכת בינה מלאכותית המשמשת כמערכת המלצה ומאפשרת קטלוג לדיים. יהיה צורך ביצוע שינויים והתקנות במערכת לטובת התאמת לדרישות הפרויקט. כגון חלופה של הלידים לאנשי המכירות לפי פיצרים והציגת האנליזות עבור סמנכ"ל המכירות (עוד מידע על המערכת ברשימת מקורות בסעיף [16](#)).

יתרונות חלופה 2:

- מערכות מתקדמות המוכינות את עצמן כבר שנים רבות במגוון תחומיים.
- תחזקה ותמייה טכנית מהספק.
- הטעמה מהירה בארגון.

חסרונות חלופה 2:

- עלויות גבוהות. על הארגון יהיה לרכוש את שירות ה-CRM עבור השימוש בתוסף של Einstein של חברת Salesforce.
- כל המידע של הארגון יctrיך לעבר המערכת CRM כיוון שיש צורך בהתממשקות מלאה וסynchron בין שתי המערכות.
- חשיפת גורם זר לנתחים של הארגון.

6.1.3 חלופה 3 - השארת המצב הקיים

בחברות הליסינג בארץ עובדים באמצעות לא טכנולוגיים - הכללים שימוש בטבלאות Excel ותקשורת בשימוש של WhatsApp. בחלופה זו לא מטבח שינוי בתהליכי הארגון. חברות הליסינג בארץ עובדות בצורה FIFO, כאשר נכנסים פרטיים חדשים ליד חדש הם מועברים על ידי סמנכ"ל המכירות באפליקציית WhatsApp לאיש המכירות הרלוונטי, והוא יוצר עמו

קשר. כל פרט הילדים מוגנים ידנית לקבצי Excel לצורך תיעוד אך לא מקבלים תיעוד מסויים או מעקב אחר הסטטוס של הלוקה הפוטנציאלי.

יתרונות חלופה 3:

- חלופה חסכונית, לא דורשת אמצעים שעולים כסף.
- אין צורך בהקמת תשתיות בשביל מערכת או שרתים מיוחדים.
- אין צורך בהכשרות עובדי הארגון עבור פעולה מערכת חדשה.
- אין סכנה של חשיפת עובדי חוץ למידע הרגיש של החברה.

חסרונות חלופה 3:

- תהליך עבודה של אנשי המכירות לא יעיל ועלול להוביל לבזבוז זמן.
- עלול לגרום פספוס לידיים או יצירת קשר לא מהירה מספיק עבור לקוחות פוטנציאלי.
- חוסר מעקב אחרי עבודות אנשי המכירות אשר עלול לגרום ל"ابتלה סמייה".
- אי מקסום רוחן החברה עקב עבודה לא עיליה ורצף עבודה לא סדיר.

6.1.4 הקритריונים להשוואה בין החלופות

6.1.4.1 התאמה לדרישות הפונקציונליות של המערכת

מהות: פרמטר זה הוא החשוב ביותר מבין הקритריונים. הקритריון בוחן את התאמה החלופות השונות לדרישות המערכת בצורה שהמערכת תקיים את כל התהליכי העבודה והפונקציונליים הנדרשים. משקל (22%): הצורך בהתאמת המערכת לדרישות הלוקה הינו גבוה ומהותי מאוד לטובות יישום התהליכי העבודה של הארגון.

דירוג: 10 - התאמה גבוהה, 0 - התאמה נמוכה.

6.1.4.2 גמישות לשינויים

מהות: פרמטר זה בוחן את היכולת לגמישות לשינויים במערכת. היכולת העתידית בכל חלופה לפתח ולשלב שינויים קטנים/aggregates במערכת הקצה. משקל (16%): הצורך בגמישות עבור שינויים עתידיים הינו גבוה ומהותי בפרויקט זה.

דירוג: 10 – גמישות גבוהה, 0 – גמישות נמוכה

6.1.4.3 הטמעה

מהות: פרמטר זה בא לבדוק את הטמעת המערכת והשימוש בה. בפרמטר זה משקללת גם כן התאמה המערכת לשיטות העבודה הקיימות, הצורך בהכשרה והדרכות, קלות הפעול ונוחות הטמעה של המערכת בארגון. משקל (15%): חווית המשמש עבור מערכת זו הינה חשובה מאוד. רק מערכת אשר מאפשר נוח ופשוט לכוכ אדם בעל הכשרה מינימלית, תוכל להשתלב בארגון ולملא את ייעודה.

דירוג: 10 – הטמעה נוחה, 0 – הטמעה לא נוחה

6.1.4.4 מורכבות הפיתוח

מהות: פרמטר זה בוחן את מורכבות הפיתוח בכל אחת מהחלופות השונות. בפרמטר זה משקללים הסיכוןים השונים בפיתוח, אשר עתידיים להשפיע על משך ועליות הפיתוח. משקל (9%): קיימת חשיבות מסוימת למורכבות הפיתוח, בעיקר בשל ההשפעה הפוטנציאלית על משך ועליות הפיתוח.

דירוג: 10 – מורכבות נמוכה, 0 – מורכבות גבוהה

6.1.4.5 משך הפיתוח וזמן מסירה לлокו

מהות: פרמטר זה בוחן את משך הזמן הצפוי לפיתוח המערכת בכל אחת מהחלופות השונות. משקל (8%): קיימת חשיבות נמוכה למשך הפיתוח כיוון וכיום הארגונים פועלם ללא המערכת.

דירוג: 10 – משך קצר, 0 – משך ארוך

6.1.4.6 עליות

מהות: פרמטר זה בוחן את העליות הצפויות לפיתוח המערכת בכל אחת מהחלופות השונות. עליות הפיתוח כוללת את הפרויקט על כל שלביו. משקל (17%): עליות הפיתוח הין חשובות מאוד. פרויקט בעלות גובהה מדי העולה על התועלת המופקת לא יכול להתבצע ותוכנתו לא תצא לפועל. בנוסף, המערכת באה ליעל תהליכי עסקים המתרחשים בארגון, אם העליות יהיו גבוהות- התועלת תהיה נמוכה.

דירוג: 10 – עלות נמוכה, 0 – עלות גבוהה

6.1.4.7 תחזקה

מהות: פרמטר זה בוחן את אחזקה הטכנית העתידית בכל אחת מהחלופות השונות. משקל (13%): קיימת חשיבות לבחינת מול האחזקה העתידי של המערכת. המערכת עתידה לפעול לפחות שנים ארוכות ויש להבטיח כי ניתן יהיה לתחזק אותה בצורה טובה ולהבטיח את פעולתה.

דירוג: 10 – תחזקה פשוטה, 0 – תחזקה מורכבת

6.1.5 השוואה בין החלופות

										משקל (%)	
		חלופה 3- פיתוח עצמי		חלופה 2- תוכנת		חלופה 1- פיתוח עצמי					
				מדד							
ציון משוקל	ציון	ציון משוקל	ציון	ציון משוקל	ציון	ציון משוקל	ציון	ציון	ציון		
0.44	2	1.54	7	1.98	9					22	התאמת לדרישות הפונקציונליות
מינימלית		אפשר לבצע התאמה		אפשר לבצע התאמה		אפשר לבצע התאמה		ריבית			
0	0	0.8	5	1.44	9					16	גמישות לשינויים
אין אפשרות לגיימות.		גמישות חילקית. לא ניתן לבצע שינויים משמעותיים בתוכנת המדד.		גמישות מלאה. המערכת תתוכנן ותבנה תוך הקמת בסיס לצרכים נוספים של הארגון							
1.35	9	0.85	4	1.2	8					15	הטעה
אין צורך בהטעה של מערכת חדשה והתפעול נוח עבור העובדים.		יהיה צורך לבצע הרבבה שינויים לטובות הטמעת המערכת הקיימת וסנכרון מלא. הקשרות עובדי החברה ייקחו הרבה זמן.		המערכת תהיה קלה להטעה וחווית המשתמש תהיה מייטבית.							
0.9	10	0.54	6	0.36	4					9	מורכבות הפיתוח
אין סיכון בחלופה זאת או מורכבות בבנייה		הפיתוח בחלופה זו הינו במורכבות בבנייה		מורכבות הפיתוח יחסית גבוהה ובעל סיכון רבים							
0.8	10	0.4	5	0.32	4					8	משר הפיתוח
אין זמן פיתוח זמן פיתוח רב											
1.7	10	0.86	4	1.19	7					17	עלויות
אין עלות עלות פיתוח בינוי		עלות פיתוח גבוהה									
1.3	10	0.52	4	0.91	7					13	תחזקה

אין צורך בתחזוקה	נוכל להשתמש בצוות התמיינית של התוכנת מדף אך נציגר עדיין להעסיק צוות טכני שմבין בקוד ובשינויים שנעשה	הפרויקט יקצה צוות טכני מלא שיוכל לתת תמיכה לארגון		
6.49	5.51	7.4		ציון

טבלה 6.1.5. השוואת בין החלופות

החלופה הנבחרת, היא חלופה 1 מערכת "יעודית לפיתוח עצמי מלא", אשר קיבלה את הציון 7.4 הינו גבוהה מבחן כל החלופות בסה"כ בכל הקритריונים שהגדכנו חשובים. כמו כן היא קיבלה גם את הציון הגבוהה ביותר מבחן הבחירה "התאמה לדרישות הפונקציונליות" שכמו שנאמר זהה הקритריון החשוב ביותר אשר עשוי להביא להצלחת הפרויקט.

6.2 תיקוף ובדיקות

6.2.1 הצגת התוצאות באמצעות ניתוחים סטטיסטיים

בחלק זה אנחנו נתיחס לניתוחים של שלושת החלקים:

- I. החלק הראשון בלמידה הבלתי מונחית כאשר אנחנו עדין לא יודעים איזה ליד הפור למכירה והמערכת משתמשת באגוריתם Prototypes – K.
- II. בחלק השני של הלמידה המונחית בה אנשי המכירות מעדכנים את הקובץ על פי המכירות שביבעו או לא, והמערכת מאמנת את המודלים ושומרת את המודל הטוב ביותר.
- III. החלק השלישי כאשר נבחר מודל הלמידה המונחה הטוב ביותר, המערכת חוזה אילו לדימוי הופיע למכירה.

6.2.1.1 חלק ראשון – ניתוח סטטיסטי של הלמידה הבלתי מונחית

בניתוח הסטטיסטי בוצע ניתוח לשולשה סטיטים של נתונים בהם השתנו במהלך הפרויקט, סט הנתונים של קובץ הילדים שהועלה על ידי הלוקוח, סט הנתונים פנימי המכיל מידע על חברות מסחריות, וסט נתונים ומידע על מכוניות.

6.2.1.1.1 ניתוח סט הנתונים של קובץ הילדים:

6.2.1.1.1.1 ניתוח המידע הנומרי:

	id_lead	id	year_of_birth	creation_time	car_year
count	1000.000000	1.000000e+03	1000.000000	1000.000000	1000.000000
mean	101500.500000	5.579014e+07	1985.988000	12.901000	2013.778000
std	288.819436	2.586079e+07	16.427468	7.210353	3.701948
min	101001.000000	1.013128e+07	1929.000000	1.000000	1998.000000
25%	101250.750000	3.370076e+07	1953.000000	6.000000	2012.000000
50%	101500.500000	5.480663e+07	1986.000000	13.000000	2014.000000
75%	101750.250000	7.753875e+07	1977.000000	20.000000	2018.000000
max	102000.000000	9.994854e+07	2003.000000	24.000000	2021.000000

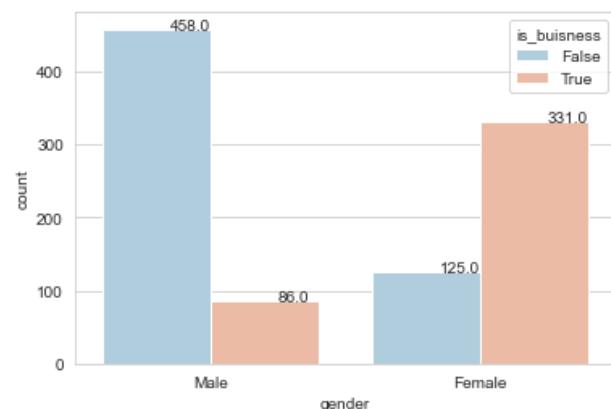
כאשר הקובץ מועלה על ידי הלוקוח ניתן להסיק על המידע הנומרי את הפרטים הבאים.

ממוצע הגילאים של הלוקוחות המתעניינים בליסינג של רכב הוא 57, האדם הכי צעיר שהתעניין בליסינג ברכב הוא ליד בן 19, והlid הכי מבוגר שהתעניין בליסינג הוא 99.

ממוצע שעوت הפניה של לידים אשר פנו לסניף הליסינג באמצעות הפלטפורמות השונות היא בשעה 00:00, והשעות הכי מוקדמות ומאוחרות הן ב-13:00 ו-21:00 בלילה.

ממוצע שנות הרכבים שבהם הלידים התעניינו להשכר בסניף הליסינג בסניף הליסינג הם רכבים משנת 1998, כאשר החzinן הוא שנת 2014, טווח שנות הרכבים הרצויים הם מ-1998 עד 2021.

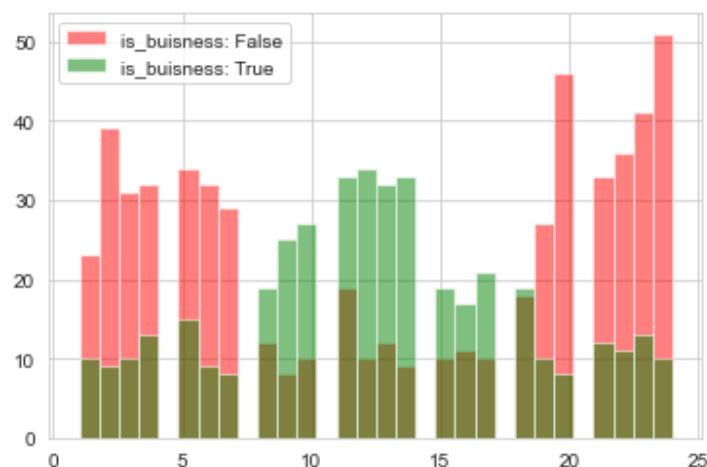
6.2.1.1.1.2 נתוח הגברים מול הנשים:



גרף 2. נתוח הגברים מול הנשים

הgraf הבא מראה שה"כ 544 גברים שהתעניינו בליסינג לעומת 456 נשים (סה"כ 1000 לדיים), אך ניתן לראות, שרוב הגברים הם למקצועות לא עסקיים לעומת נשים שהובן למקצועות עסקיים שהתעניינו בליסינג לצורך הארגון בהן עוסקות.

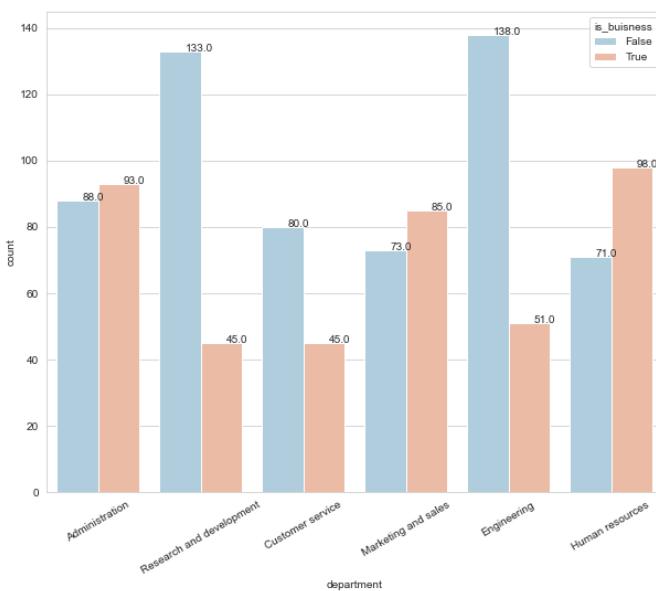
6.2.1.1.1.3 נתוח שעות הפנינה של הליד:



גרף 3. נתוח שעות הפנינה של הליד

התרשימים הבא מציג שרוב הלמקצועות עסקיים (ירוק) לעומת נסנייף הליסינג באמצעות הפלטפורמות השונות בשעות הבוקר-צהרים, לעומת זאת שרוב הלמקצועות לא עסקיים (אדום) ניתן לראות שרוב הלמקצועות העסקיים פנו לסנייף הליסינג בשעות הערב המאוחרות או הבקර המקודמות.

6.2.1.1.1.4 ניתוח המחלקה בה עובד הlid:



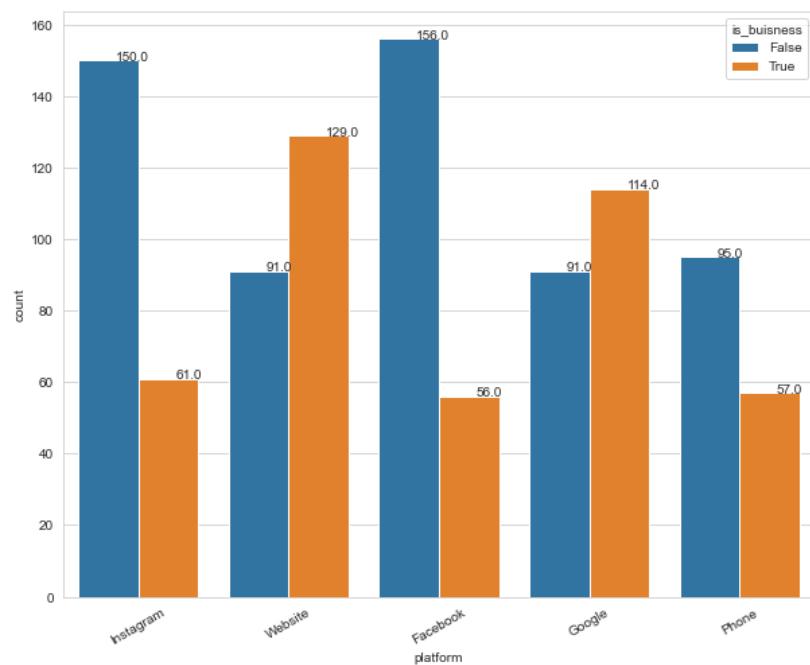
גרף 6.2.1.1.1.4 ניתוח המחלקה בה עובד הlid

ניתן לראות באמצעות התרשימים שרוב הלוקוחות הלא עסקיים עובדים במחלקות של: מחקר ופיתוח, שירות לקוחות, מהנדסים.

לעומת הלוקוחות העסקיים שעובדים במחלקות שיווק ומכירה, משאבי אנוש, ואדמיניסטרציה למוראות שיש הבדל מאד קטן לעומת הלוקוחות הלא עסקיים.

מסקנה זו מתאיימה למידע שנמסר לנו מחברות הליסינג שאיתן ביצעו את סקרי השוק, שלוקוחות עסקיים הפונים לחברת ליסינג לצורך עבודה בדרך כלל עוסקים במחלקות של שיווק ומכירה ומשאבי אנוש, לעומת מהנדסים ונותני שירות, שפונים לחברת הליסינג לצורך עצמי.

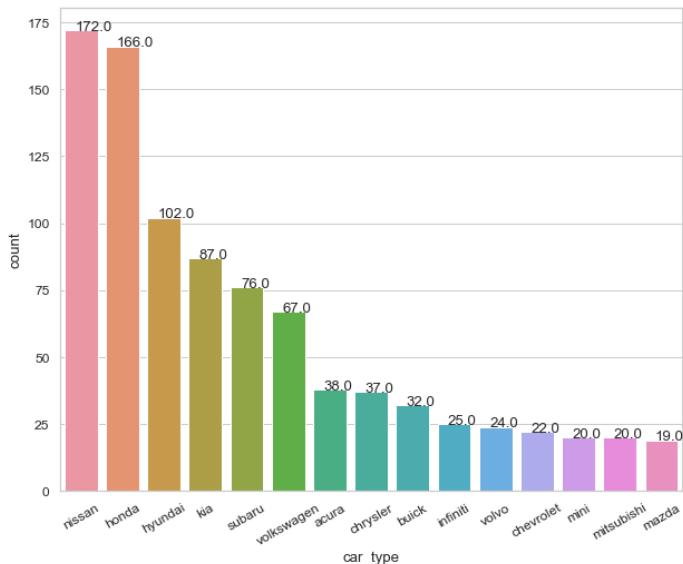
6.2.1.1.1.5 ניתוח הפלטפורמה ממנה הגיע הlid



גרף 6.2.1.1.1.5 ניתוח הפלטפורמה ממנה הגיע הlid

לפי התרשימים הבא ניתן לראות שרוב הליקוחות הלא עסקיים פנו אליו בפלטפורמות: אינסטגרם, פייסבוק ודרכ הטלפון, לעומת זאת הליקוחות עסקיים שפנו אליו דרך האתר, ודרך גוגל. חברת הליסינג מבצעת שיווק דיגיטלי בכל הפלטפורמות הללו ושמורת את המידע מאיזה "אטר נחיתה" ניגש אליו הlid.

6.2.1.1.6 15 חברות הרכבים הכיכי מובוקשות על ידי הלידים:



6.2.1.1.6 15 חברות הרכבים הכיכי מובוקשות על ידי הלידים

ניתן לראות באמצעות התרשימים הבא שרוב הלידים התעניינו ברכבים ממחברות של ניסן, הונדה והיונדאי.

עובדה זו מראה על אופי הליקוחות והמכוניות בהן הם מתעניינים, ונitin להסיק שרוב הליקוחות מתעניינים ברכבים סטנדרטיים ולא ברכבי יוקרה.

6.2.1.1.2 נתוח סט הנתונים של קובץ הרכבים:

בידינו סט נתונים של רכבים אשר נאוסף לצורך הוספה מידע גומרי וסטטיסטי על הרכבים שבו השתמש בשbill לנתוח את מחיר הרכב.

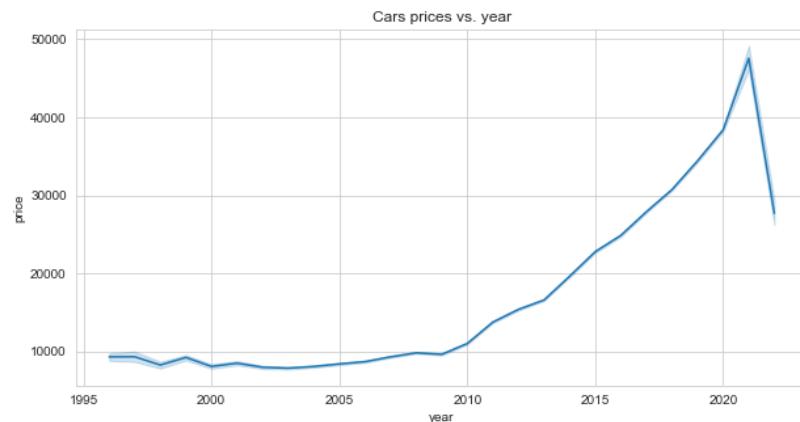
6.2.1.1.2.1 נתוח המידע הnumerico:

	id_car	price	year
count	3.433880e+05	343388.000000	343388.000000
mean	7.311433e+09	20094.093041	2012.767557
std	4.385956e+06	14616.009126	5.262639
min	7.301583e+09	2488.000000	1996.000000
25%	7.308030e+09	8500.000000	2009.000000
50%	7.312442e+09	16900.000000	2014.000000
75%	7.315190e+09	28590.000000	2017.000000
max	7.317101e+09	347999.000000	2022.000000

ממוצע מחירי הרכבים בסט נתונים הרכבים שבידינו הוא K 20 דולר, מחיר הרכב הכיכי זול הוא 2488 דולר, והרכב הכיכי יקר הוא K 348K דולר.

ממוצע שנת הרכבים הוא 2012 כאשר יש לנו מידע על רכבים יישנים משנת 1996, ומידע על רכבים חדשים משנת 2022.

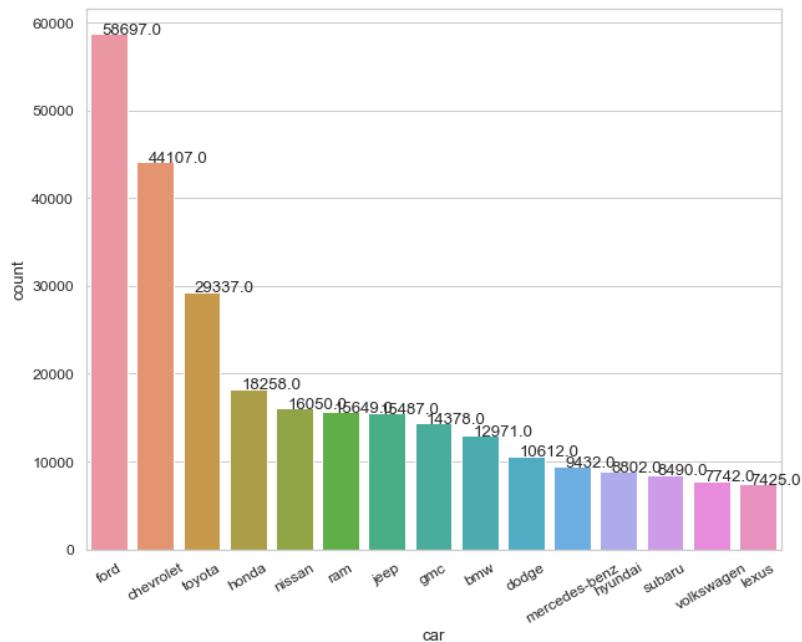
6.2.1.2.2 מחיר הרכב לעומת השנה שלו:



גרף 6.2.1.2.2. מחיר הרכב לעומת השנה שלו

ניתן לראות שיש קשר ליניארי מובהק בין מחיר הרכב לעומת השנה שלו, אך לאחר שנת 2021 מחיר הרכב יורד משמעותית, את המידע של סט הנתונים אספנו מאגרים שונים באינטרנט, והיה לנו מידע לוקה בחסר על רכבים יקרים במיוחד משנת 2022 מכיוון שהיו חדשים מדי (אך אין לכך השפעה על המודל והסיווג).

6.2.1.2.3 15 חברות יצרני המכוניות הכי שכיחות:



גרף 15. 15 חברות המכוניות הכי שכיחות

ניתן לראות שבבדטה סט של המכוניות יש הכי הרבה מידע על רכבים מחברת פורד ושרבולט, אך מידע זה אינו הכרחי לתהילר הסיווג מכיוון שאנו צריכים במידע זה לצורך הצלבת הנתונים עם המכוניות שהילדים התעניינו לגבייהם וב_udרתו נחשב להם את מחיר הרכב המבוקש.

6.2.1.1.3 ניתוח סט הנתונים של קובץ החברות העסקיות

6.2.1.1.3.1 ניתוח מידע המומרי

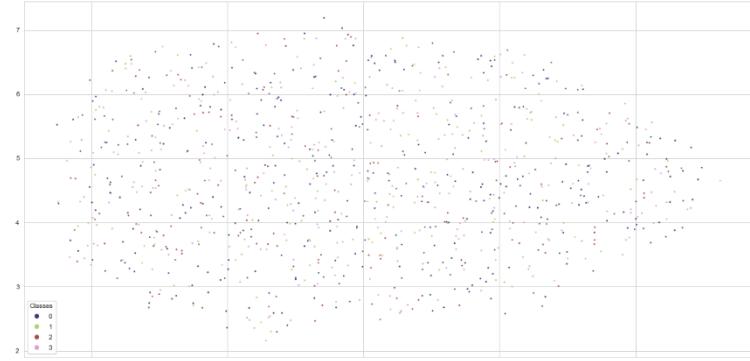
	id_company	rank	rank_change	revenue	profit	num. of employees
count	997.000000	997.000000	997.000000	997.000000	997.000000	9.970000e+02
mean	30498.000000	500.144433	0.437312	15936.748546	1345.940120	3.467420e+04
std	287.953411	289.122799	22.455474	34809.536601	4517.936544	9.215434e+04
min	30000.000000	1.000000	-186.000000	1990.300000	-8506.000000	5.100000e+01
25%	30249.000000	250.000000	0.000000	3163.800000	110.500000	6.400000e+03
50%	30498.000000	500.000000	0.000000	5655.000000	380.900000	1.300000e+04
75%	30747.000000	751.000000	0.000000	12856.000000	1062.000000	2.905800e+04
max	30996.000000	1000.000000	224.000000	523964.000000	81417.000000	2.200000e+06

עמודת הדירוג מבטא את הדירוג של החברות בראשימת החברות המובילות בשוק, لكن היא עמודה סדרתית אשר לא נתנת לנו ערך נוסף, לעומת זאת עמודת Rank Change היא שינוי דירוג החברה משנה בעברה.

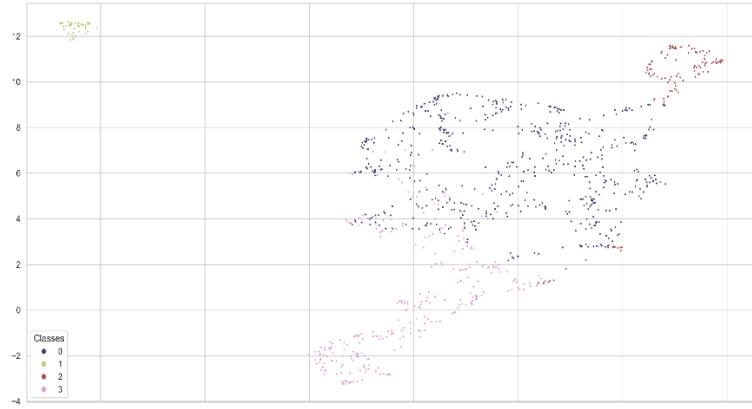
ממוצע ההכנסות של החברות הוא 15,936 דולר הסכום הנמוך ביותר הוא 1990 דולר, והגבוה ביותר הוא 522K דולר.

ממוצע הרוחחים של החברות הוא 1345 דולר, כאשר הרוח הגדול ביותר הוא 814K דולר, והנמוך ביותר הוא 8K.

6.2.1.1.4 תוצאות החלוקת לאחר הפעלת האלגוריתם בצורה ויזואלית בתחילת הפרויקט, תוצאות החלוקת שלנו בצורה ויזואלית נראה כך:



אלה הן תוצאות החלוקת העדכנית של האלגוריתם:



ניתן לראות בבירור שיפור משמעותו בעבודת האלגוריתם בחלוקת לקבוצות, וזאת כתוצאה מלמידת שוק הרכבים והליסינג ומכך שיפורנו את הדטה סט של הלידים.

לפי התרשימים הבא, ניתן לראות שיש 4 קבוצות איחודות כאשר כל קבוצה מייצגת קבוצה מסוימת של לדיים, לדיים "לוהטים", לדיים "", לדיים "ביןניים" ו-לדיים "קרים".

לפי הציגה הויזואלית, נראה שהאלגוריתם הצליח לחלק לקבוצות אלו בצורה טובה.

מה שתומך במידד ההצלחה שקבענו לעצמנו בסעיף [2.4.1](#)

Silhouette Score 6.2.1.1.5

מידע על Silhouette Score בסקירת הספרות [3.6.1](#)

הציון שקיבלנו על פי מידד ה-Silhouette הוא 0.27

Silhouette Score

```
In [241]: round(silhouette_score(numerical, clusters),2)
```

```
Out[241]: 0.27
```

כאשר טווח הציונים הוא בין -1 ל 1, ציון זה נחשב טוב, אך עדין טוון שיפור.

הסיבה שקיבלנו את ציון זה נובעת מכך שביצעו נרמול על הנתונים המ名义ים בשביל שהאלגוריתם יכול לחלק את הלידים בצורה יותר יותר וימנע מלתת משקל יתר לערכים קיינוניים.

בעקבות הנרמול הנתונים יותר דומים, ובכך המטריקה נותנת ציון נמוך יותר כי ההבדל בין הנקודות מצטמצם.

בנוסף, ביצוע הסילווט אין התייחסות לעמדות הקטגוריאליות שמהוות חלק חשוב באפיון תוצאות החלקה.

לכן למרות ציון זה, אנחנו שבעי רצון מחלוקת הקבוצות של האלגוריתם.

6.2.1.1.6 תוצאות החלקה של האלגוריתם הבלטי מונחה

	Segment	Total	Is_buleness	gender	department	car_year	platform	age	car_price	desirable_rental_days	time_catagor	Market Cap	profit
0	First	530	False	Male	Engineering	2015.0	Instagram	61.5	13003.130047	232.0	Night	9243.209434	410.628868
1	Second	41	True	Female	Marketing and sales	2020.0	Google	40.0	107812.565854	607.0	Morning	401741.024390	22527.634146
2	Third	123	False	Male	Research and development	2007.0	Facebook	54.0	7195.731789	242.0	Evening	13439.900000	-168.339024
3	Forth	306	True	Female	Human resources	2015.0	Website	48.0	14675.332876	544.0	Noon	14547.377124	628.356209

על פי הטבלה הבאה, ניתן לראות את תוצאות החלקה של האלגוריתם, תוצאות החלקה מאפיינות את העמדות שהיו כדי חשובות באפיון הליד, ואלה שהתקבלו כקלט באלגוריתם.

הקבוצה הדומיננטית ביותר, הינה קבוצה מספר שתיים המכילה 41 לדיים.

רוב הקבוצה הזה הינם לדיים עסק'ים, המין הנפוץ ביותרן הן נשים שלחוב עובדות במחלקות של שיווק ומכרה, שנת הרכבת המבוקשת ביותרן הם רכבים 2020, הפלטפורמה הנפוצה ביותרן ממנה הגיעו הלידים היא גוגל, ממוצע הגלאים הוא 40, מכיר הרכבת הממוצע הוא 100K, כמהות הימים הממוצע שמדובר להשכרה הוא 607, לרובם פנו בשעות הבוקר, והחברות מבוחנת שווי שוק ורווח הוא הגבוהה ביותר, لكن ניתן להסביר קבוצת הלידים הזה הם לדיים "רווחחים" שיש להם סיכוי גבוה להפוך למכירה או סיכוי להניב את הרוחם הגבוהה ביותר לסתף.

לעומת זאת קבוצה מספר 3 המכילה 123 ילדים, רובם הם ל��וחות פרטיים, המין הפופולרי ביותר הם גברים, המחלקה השכיחה ביותר היא מחקר ופיתוח, ממוצע שנת הרכב המבוקש הם רכבים מ-2007, הפלטפורמה ממנה נחתו הילדים היא פיסבוק, מחיר הרכב הממוצע הוא הנמוך ביותר - 7K, הילדים מעוניינים בהשכלה לתוויה הקצר ביותר, לרבות פנו בשעות הערב, ושווי השוק והרווחים הנמוכים ביותר, لكن ניתן להסיק שקבוצת הילדים זו הם "קרים".

קבוצה מספר ארבעה הם 306 ילדים מהם לרוב ל��וחות עסקיים, המין הנפוץ ביותר הן נשים, שעובדות לרוב במחלקות משאבי אנוש, שנת הרכב הממוצעת בה הילדים מעוניינים הם רכבים מ-2015, שפנו אליו דרך האינטרנט, הגיל ממוצע הוא 48, מחיר הרכב הממוצע הוא 14K, כמות הימים שעוניינים בה להשכלה הוא השני לאחרו, שלרוב פנו אליו בצהרים, ורוחוי החברה ושווי השוק הם השניים הגדולים ביותר, אך ניתן להסיק שנייה לסוג את קבוצת הילדים זו כילדים "חמים".

קבוצה מספר אחת הקבוצה הגדולה ביותר המכילה כ-534 ילדים, הם לרוב ל��וחות לא עסקיים, המין הפופולרי ביותר הם גברים, המחלקה הפופולרית ביותר בה עובדים בחברה שלהם היא הנדסה (למרות שהם עובדים גם במחלקות של משאבי אנוש ושירות לקוחות), שנת הרכב הממוצעת הם רכבים מ-2015, פלטפורמת הנחיתה היא אינסטגרם, הם אנשים ממוצע בגילאי 61, מחירי הרכב שבהם הם התענינו בממוצע 12.9K, כמות הימים בהם מעוניינים בה להשכלה הם 233, פנו אליו לרוב בלילה וערב, והם עובדים בחברות שהרווחים שלהם ושוקי השוק הם ממוצעים.

לכן ניתן לסוג את ילדים אלה כילדים "בינוניים".

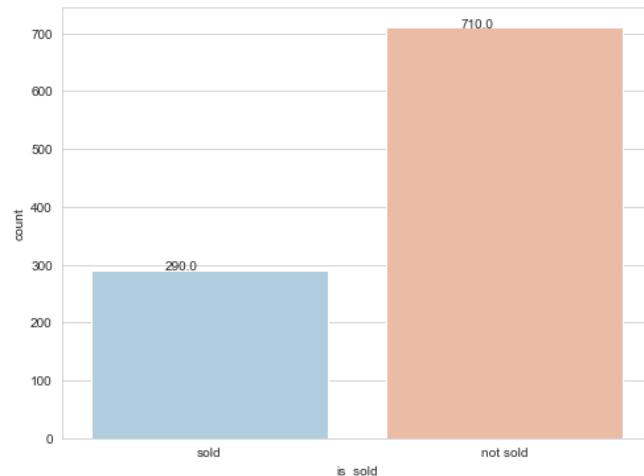
6.2.1.2.3. המידע כיצד יוצרים את הדאטה פריים והגענו לתוכאות הסיכום של כל אשכול בסעיף

6.2.1.2.3.1. חלק שני – ניתוח סטטיסטי של אימון המודלים

בחלק זה אנחנו נבחן את נתוני הילדים לאחר שנאנו המכירות ביצעו ניסיונות מכירה על קובץ המקוטלג, וכעת יש לנו את עמודת `sold`_זאת המצינית איזה ליד הפר למכירה.

ナルם אילו ילדים הפכו למכירה ומה המסקנות שעולות מכך.

6.2.1.2.3.1.1. כמות מכירות שנסגרו

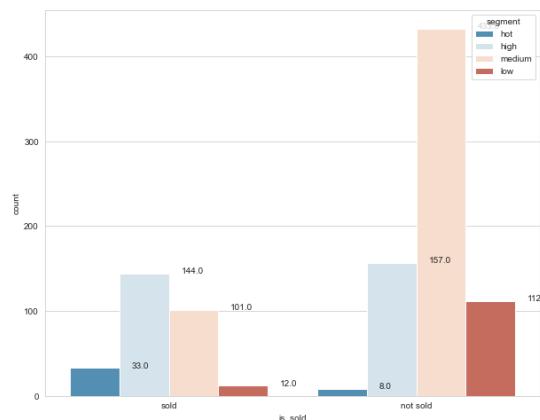


גרף 6.2.1.2.3.1.1. כמות המכירות שנסגרו

על פי תרשيم זה ניתן לראות ש-290 ילדים נמכרו לעומת 710 שלא נמכרו, סה"כ סניף הליסינג הצליח למכור כמעט 30% מהילדים.

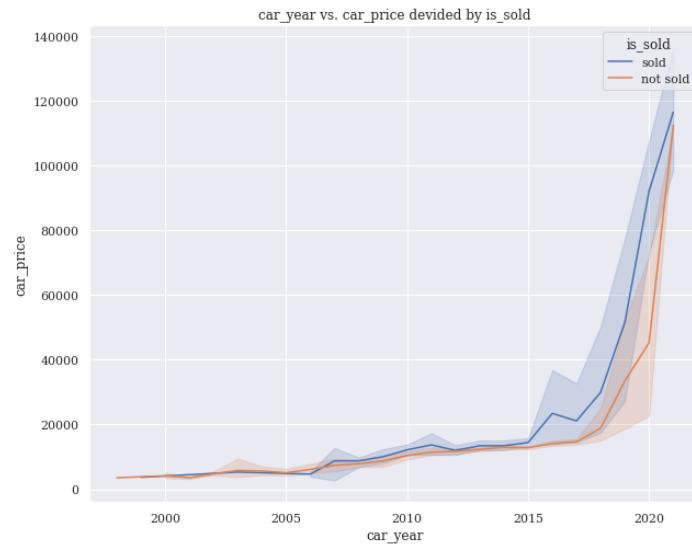
עמדנו במידד הצלחה של הפרויקט [2.3.1](#) שהוא הפיכת 30% מהילדים החמים אשר הפכו למכירה.

6.2.1.2.2 כמות מכירות שנסגרו לעומת חלוקה שביצעו בחלק הראשון



גרף 6.2.1.2.2. כמות מכירות שנסגרו לעומת חלוקה שביצעו בחלק הראשון
 גраф זה חשוב מאוד ומשקף את המדרדים אליום התching'בנו בתחילת הפרויקט, מدد הצלחה שנקבע הוא הפיכת 30% מהילדים החמים למכירה [2.3.1](#).
 ניתן לראות ש-33 מהילדים אשר קיבלו את הסיווג "lohutim" נמכרו, לעומת 8 ילדים שלא נמכרו, כלומר סה"כ 80% מהילדים הlohutim הפקו למכירה.
 הסניף הצליח למcor 50% מהילדים החמים, 19% מהילדים הבינווניים וכ-9% מהילדים הקרים.
 לכן ניתן להגיד שהסיווג הבלתי מונחשה שהאלגוריתם שלמו ביצע בחלק הראשון תואם מבחינה סטטיסטית למכירות שהתבצעו בפועל.

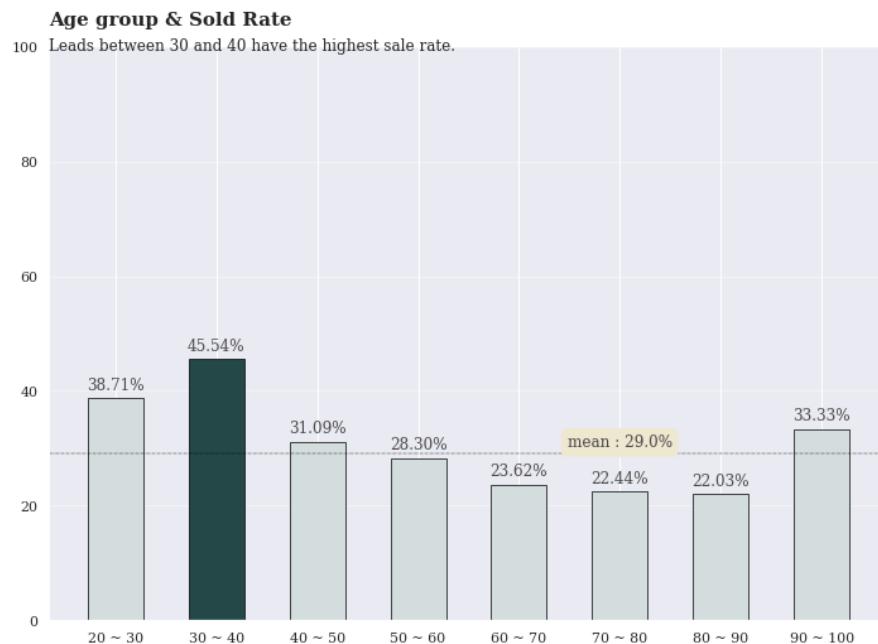
6.2.1.2.3 שנת הרכב מול מחיר הרכב – האם נמכר או לא



גרף 6.2.1.2.3. שנת הרכב מול מחיר הרכב – האם נמכר או לא
 ניתוח הסטטיסטי הנ"ל ניתן לראות את שנת הרכב בציר ה-X מחיר הרכב בציר ה-Y, והקו הכהול מתאר הילדים שנמכרו והקו הכתום מתאר ילדים שלא נמכרו.
 ניתן לראות שבהתחלת המכירות מתפלגות אותו דבר, אך החל מ-2015, סניף הליסינג באוטו מחר הצליח למcor רכבים יפנים יותר לעומת רכבים חדשים שבהם הילדים התעניינו, שלבסוף לא הפקו למכירה.

ניתן גם לראות שמחיר הרכב שמחיר הרכב הכי יקר שהושכר יותר גדול ממחיר הרכב הכי יקר שלא הושכר, שכן ניתן להסיק שלרוב האנשים שהתעניינו בהשכרת רכבים יותר זולים יש קורלציה לכך שהם לא הפקו למכירה.

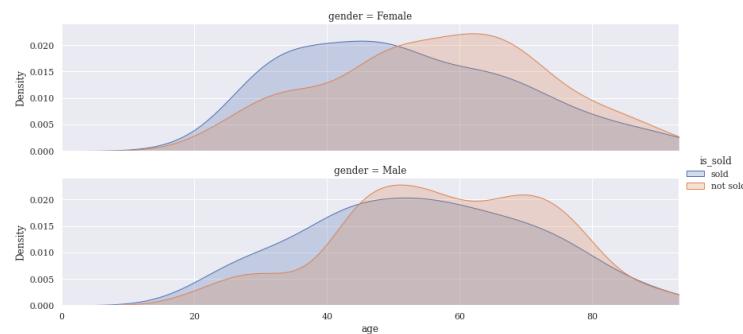
6.2.1.2.4 קבוצת הגילאים שהפכו למכירה



גרף 6.2.1.2.4 קבוצת הגילאים שהפכו למכירה

ניתן לראות על פי הגרף איזה איזה קבוצת גילאים הפכה למכירה ומה ערך שיעור ההמרה למכירה. ניתן לראות שקבוצת הגילאים הכי פופולרית שהפכו למכירה היא קבוצת אנשים מגילאי 30 עד 40, וערך המריה למכירה שלהם הוא, 45.54%, לעומת זאת חצי מהילדים בקבוצת הגילאים זו הפכו למכירה, נתון זה הגיוני כי לחוב האנשים שמתעניינים בלייטינג נמצאים בקבוצת גילאים זו. הקבוצה השנייה הם אנשים בגילאי 20 עד 30, ואחריה גילאי 90 עד 100, נתון שנותן זה מצין את אחוז ההמרה, סה"כ יש בסט הנתונים רק 15 לקוחות מעל גיל 90, מתוך כולם 5 הפכו למכירה שכן קיבלו שליש מהם נמכרו.

6.2.1.2.5 התפלגות צפיפות המכירות גברים מול נשים



גרף 6.2.1.2.5 התפלגות צפיפות המכירות גברים מול נשים

על פי ניתוח זה, ניתן לירות 2 תרשימים כאשר ציר ה-Y זה צפיפות המדגם, ציר ה-X זה גיל הגיל, הגרף העליון הוא מדגם הנשים והתחתון של הגברים, צבע הכתול האם הלידים נמכרו, צבע אדום לא נמכרו.

ניתן לראות שהגיל ב-2 המינים מתפלג נורמלי לעומת המכירות, כאשר רוב המכירות התרכזו בגילאי עד 55 אצל הנשים.

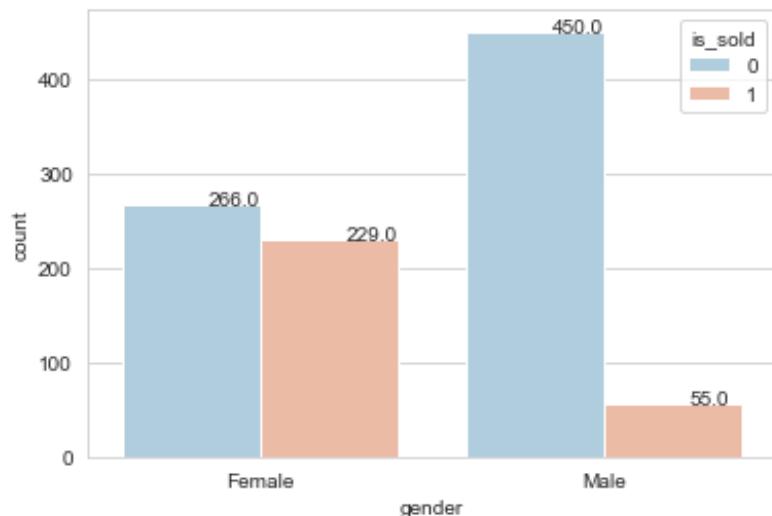
גברים היה פחות הבדל ועקומת המכירה לפי הגילאים יותר שטוחה, כלומר, היו הרבה יותר לברים גם בגיל יותר מאוחר שהפכו למכירה לעומת הנשים.

באופן כללי ניתן לראות שצפיפות המכירה אצל הנשים יותר גדולה מאשר הגברים, כלומר יש יותר נשים שהפכו למכירה מאשר גברים (אך לא בהבדל משמעותי).

6.2.1.3 חלק שלישי – ניתוח סטטיסטי לאחר חיזוי אילו לידים הפכו למכירה

בחלק זה נבצע ניתוח סטטיסטי לאחר שהלכנוعلاה את הקובץ בפעם השלישייה, קובץ חדש אשר המערכת לא למדה בעבר, לאחר תהליך preprocessing, נבצע חיזוי על הקובץ ונראה אילו לידים הפכו למכירה.

6.2.1.3.1 ניתוח מכירות הגברים מול הנשים

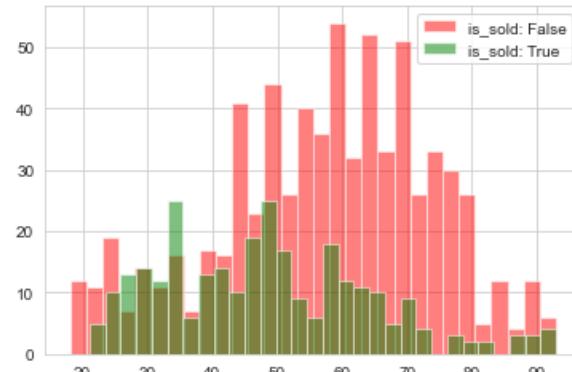


גרף 6.2.1.3.1 ניתוח מכירות הגברים מול הנשים

ניתן לראות על פי התרשימים, שכמו שקרה בפועל, יש הבדל משמעותי בין כמות הנשים שהוא חזה שהפכו למכירה לעומת הגברים, שהוא חזה שרבים לא יהפכו למכירה.

סה"כ המודל חזה ש-284 לידים יהפכו למכירה, מתוכם כ-229 נשים, כלומר הוא חזה ש-80% מהנשים בדאטא סט החדש יהפכו למכירה.

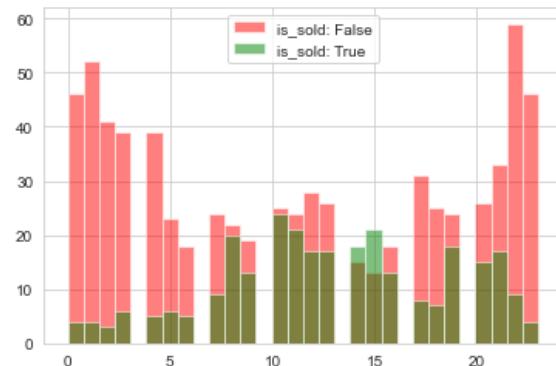
6.2.1.3.2 גילאי האנשים שהפכו למכירה לעומת כאלה שלא



גרף 6.2.1.3.2 גילאי האנשים שהפכו למכירה לעומת כאלה שלא

ניתן לראות שבדומה לדאטה סט המקורי, המודל חזה בדאטה סט החדש שרוב הלידים שהפכו למכירה הם בגילאי 35 עד 50, כאשר הוא נתן חשיבות מסוימת גם לאנשים מבוגרים. אך ניתן לראות רוב של אנשים מעל גיל 60, או אנשים בסביבות גיל ה-20 שרואים בברור שהם לא הפכו למכירה, מה שתואם לדאטה המקורי.

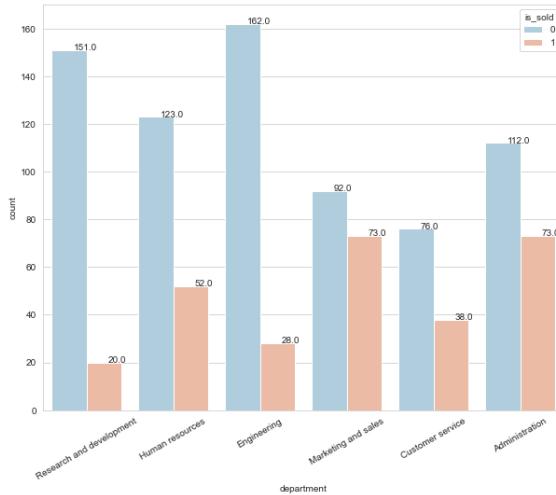
6.2.1.3.3 ניתוח שעת פניה הליד לעומת האם



גרף 6.2.1.3.3 ניתוח שעת פניה הליד לעומת האם

בתרשים זה ניתן לראות שהמודל שלנו חזה שרוב הלידים שנולדו אלינו בבוקר-צוהרים הפכו למכירה, נזכיר שבסעיף [6.2.1.1.7](#), ראיינו בתוצאות החלוקה שלידים שהוגדרו "ЛОחותים" הם לידים שנולדו אלינו בבוקר – צוהרים, لكن תחזית זו מחזקת את טיב של הלמידה הבלתי מונחית.

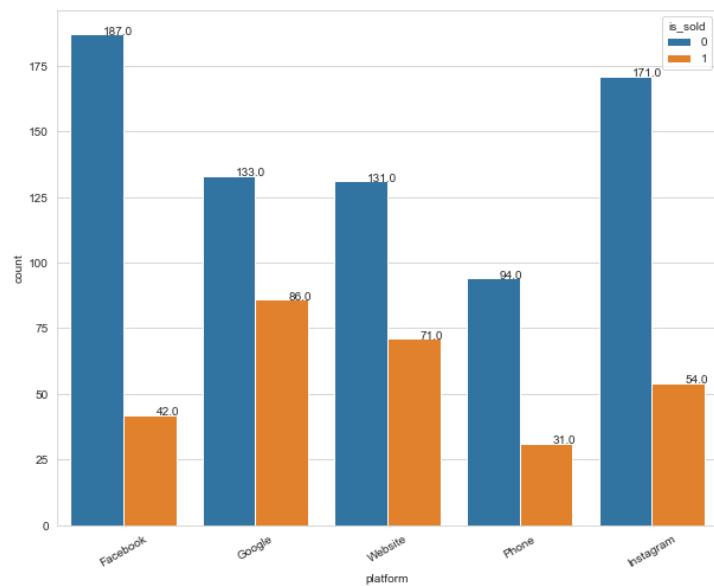
6.2.1.3.4 ניתוח מחלוקת שבת עובד הליד לעומת המכירה או לא



גרף 6.2.1.3.4 ניתוח מחלוקת שבת עובד הליד לעומת המכירה או לא על פי תרשימים הסכימה ניתן לראות שרוב הלידים שהמודל חזה שהם לא הפכו למכירה הם לידים אשר עובדים במחלקות של מחקר ופיתוח, והנדסה, ניתן לראות שלמרות שבADMINISTRATIVE ו-SALES המכירה הרוב נחזה שלא הפכו למכירה, עדין היחס בין האם המודל חזה שהם הפכו למכירה או לא יהיה יותר מאוזן משאר הקטגוריות.

חיזוי זה גם מתאים לדאטה סט המקורי, אשר רוב הלידים שעובדים במחקר ופיתוח הנדסה ושירות לקוחות הם בדרך כלל לקוחות לא עסקים שיש להם סיכוי נמוך יותר להפוך למכירה.

6.2.1.3.5 ניתוח פלטפורמות המכירה לעומת למכירה



גרף 6.2.1.3.5 ניתוח פלטפורמות המכירה לעומת למכירה מראה לנו את היחסים בין הפלטפורמות שהפכו למכירה לבין הפלטפורמות שהפכו לאuction. ניתן לראות שהתפלטפורמות שהפכו למכירה (Facebook, Google, Website) רוב הליקותיהם נזקקן למכירה, בעודם שהפכו לאuction. אינסטגרם, פיסבוק וויליאמסון מילויים מילויים למכירה. ככלומר, רוב הליקותיהם נזקקן למכירה.

נתון זה מוכיח לנו את ההשערות שהנחנו כשבנו את המודל הבלתי מנוחה, כאשר ההנחה הייתה שרוב הליקותם של האנשים מילויים מילויים כמו פיסבוק, אינסטגרם בדרכם כלל הם ליקותם מילויים מילויים "רציניים" אשר מעוניינים ביליאנס לשימוש עצמי, וכך היסכוי שלהם להפוך למכירה מילויים מילויים.

6.2.1.3.6 תוצאות החיזוי של האלגוריתם המונחה הטוב ביותר ביחס

Result	Total	is_buisness	gender	department	car_year	platform	age	car_price	desirable_rental_days	Market Cap	profit
0	0	716	False	Male	Engineering	2013.0	Facebook	56.923184	11590.724120	248.5	10894.927793
1	1	284	True	Female	Marketing and sales	2016.0	Google	48.151408	35231.386514	517.5	106025.614085

על אופי חישוב תוצאות החלוקת של האלגוריתם והאלגוריתם הטוב ביותר ביחס לתוצאות השבחה נפרט בסעיף 11. בחלק זה רק ננתה את התוצאות הסטטיסטיות של האלגוריתם המונחה הטוב ביותר. ניתן לראות שהוא ש-716 לא יפה למכירה לעומת 284, נזכיר שההתוצאות האמיטיותicas אשר אנשי המכירות ניסו למכור את הלידים המקוטלים, הם הצליחו למכור 290 ולא הצליחו 710. ככלומר, גם החיזוי של מספר המכירות תואם ברמה גבוהה לתוצאות האמיטיות בדעתה סט של החלוקת.

רוב הלידים שלא נמכרו הם לידיים שאינם ליקות לא Uskym, המין השכיח ביותר הם גברים, המחלוקת השכיחה ביותר היא הנדסה (למרות שיש גם מחלוקת אחרת שהם הוא סיווג שלא בוצע מכירה), שנת הרכבת הממוצעת הם רכבים מ-2013, פלטפורמת הנחיתה הפופולרית היא פיסבוק, גיל הממוצע הוא 57, תקופת ההשכרה הרציפה היא 248 ימים, שוקי השוק של החברה בהם הלידים עובדים הוא 10K, והרווח של החברות הוא 325.

לעומת זאת, הלידים שהוא סיווג שיפה הם לרוב ליקות Uskym, אשר המין השכיח ביותר הן נשים, שעבודות במחלוקת של שיווק וממכירה, שנת הרכבת הממוצעת הם רכבים מ-2016,

פלטפורמת הנוחיתה הפופולרית היא גוגל, הגיל הממוצע הוא 48, מחיר הרכב הממוצע הוא K.35K. תקופת ההשכלה הממוצעת הרצiosa היא 517 ימים שווי השוק הוא K.106K, ורווח החברה הוא K.4.8K.

סיכום: ניתן לראות דמיון גדול בין תוצאות החלוקה של האלגוריתם הבלטי מונחה בסעיף 6.2.1.1.7 לעומת תוצאות החלוקה של האלגוריתם המונחה.

הוא חזה שייפכו למכירה לרוב לקוחות עסקיים אשר מעוניינים להשכרת ליסינג לטוווח הרחוק, בד"כ הם אנשים יותר צעירים אשר עובדים בחברות יותר מרוחיקות בחלוקת של מכירה ושיווק.

תוצאות אלה אינטואטיביות ועוננות לנו גם על הסקירות ספורות והחקירה שביצעו על עולם הלידים של הרכבים, וגם על המטרה העיקרית של הפרויקט, בניית מערכת המלצה חכמה שתאפשר לאנשי המכירות לנוהל לידים פוטנציאליים בצורה יילה ולחזות אילו לידים בעלי הסתברות גבוהה ביותר ביותר להפוך למכירה.

6.2.2 התוצאות שהתקבלו מול אלו שנקבעו

לאחר סקירה עמוקה של התוצאות שהתקבלו במלידה המונחת ובמלידה הבלטי מונחת, ולאחר הצגת אב טיפוס המערכת (האתר), ניתן לקבוע שהצלחנו לעמוד ברוב מטרות יעדים ומדדים שנקבעו כדי הצלחה.

6.2.2.1 התוצאות למטרת הפרויקט

מטרת הפרויקט במסמך ה-SOW הייתה: בניית מערכת המלצה חכמה שתאפשר לאנשי מכירות לנוהל לידים פוטנציאליים בצורה יילה תוך מתן תמונת מצב של המכירות בכל זמן עבור הדרוג הבכיר).

אכן בנוינו מערכת המלצה שנوتנתה חיזוי מהיר לאנשי המכירות אילו לידים בעלי הסתברות גבוהה להפוך למכירה.

במהלך הפרויקט שינוינו את מטרת הפרויקט והחליטנו לא למש את תמונה המצב בכל זמן נתון.

מכיוון שהבנו שלא אפשר מבחינה פונקציונליות של המערכת לאפשר קבלת מידע ב"סטרימינג" (כלומר בכל רגע נתון לקבל מידע חדש) ולנתח אותו באופן מיידי, החלטנו שהמטרה שלנו היא לנתח ולסואג לידים באופן מיידי ולחזות אילו לידים ייפכו למכירה, ובנוסף להציג את מצב המכירות בדASHBOARD "יעוד".

המידיות מתבצעת על ידי העלאת הראשונה של קובץ הלידים, כאשר על ידי אלגוריתם למידה בלטי מונחה אנחנו מחלקים את הלידים לקבוצות, ולאחר מכן על ידי אלגוריתם מונחה אנחנו חוזים אילו לידים ייפכו למכירה.

6.2.2.2 התוצאות למדדי ההצלחה שנקבעו

התוצאות למדדי ההצלחה שנקבעו בוצע בסעיף 2.5 אשר שם פירטנו על כל מדד האם הפרויקט הצלח בשלב זה לעמוד בו ומסקנות להמשך.

7 אפיון המערכת

7.1 סיכום תהליכי כריית המידע, ETL וה-Pre-processing

בחלק זה, אנו נחלק את סיכום תהליכי כריית המידע לשלושה חלקים:

- I. בחלק הראשון כאשר הלוקה מעלה את קובץ הלידים בפעם הראשונה, מופעל אלגוריתם משפחתי ה-*Unsupervised Learning* (*K-Prototypes Algorithm*).
- II. בחלק השני כאשר הלוקה השתמש בקובץ הלידים המוחולק ואנשי המכירות ביצעו מכירות מופעל אלגוריתמים משפחתיים ה-*Supervised Learning*.
- III. בחלק השלישי כאשר הלוקה מעלה קובץ לידים חדש לאחר שהמערכת למדה אילו לדיים הפכו למכירה אצל הלוקה, המערכת משתמשת במודל הטוב ביותר שלמדה בשלב השני, ובמצעת סיווג על קובץ הלידים החדש.

7.1.1 חלק ראשון – שימוש באlgorigthm הבלטי מונחה (הלוקה מעלה את קובץ הלידים בפעם הראשונה)

(קישור לקובץ PDF המציג מחברת מסוג Jupyter Notebook שמתאר את תהליכי כריית המידע של החלק הראשון בסעיף [17.6](#)).

7.1.1.1 יבוא הנתונים

יש לנו שלושה מקורות נתונים, מקור הנתונים העיקרי הוא קובץ הלידים מסוג CSV המכיל פרטים על לקוחות שהביעו התעניינויות בפלטפורמות השונות בסגירת חוזה ליסינג, ניתן לראות את קובץ הלידים בסעיף [17.6](#), ואת מילון הנתונים [8.5.5.1](#).

הנתונים הללו מכילים פרטים כמו: שם פרטי, שם משפחה, מין, תאריך לידיה, רכב מודף, האם הלוקה עסק ועד פרטים שונים שבאמצעותם נוכל להפעיל אלגוריתמים של מידת מכונה ולחזות אילו לקוחות בהסתברות גבוהה יהיו מעוניינים בסגירת חוזה ליסינג.

בפרויקט שלנו אנחנו מודמים סניף של חברת ליסינג, הנתונים שלנו מtabsets על פי סקר הספרות שערכנו וחיפוש אינטרנט על מה קובץ לדיים של חברת ליסינג אמר או להכיל. בשבייל להכין את קובץ הלידים השתמשנו ב-*Data Generator* בשבייל לדמות קובץ זה ונשתמש בו בשבייל להוכיח את טיב המערכת שלנו ושהיא עובדת בצורה טובה ויכולת למיין את הלידים האלו לлокחות.

מקור הנתונים השני הוא מאגר מידע על רכבים המכיל פרטים כמו סוג רכב, שנת הרכבת, מחיר הרכב, מודל הרכב ועוד.

ניתן לראות את מאגר הרכבים בסעיף [17.6](#) ואת מילון הנתונים [8.5.5.3](#).

אנו משתמשים בקובץ זה במהלך הפרויקט על מנת לחשב את מחיר המחרון המעודכן של הרכבים בהם הלידים הביעו התעניינויות, ולמחיר הרכב יש משקל קבוע כדי לחזות אילו לדיים בהתאם למחיר המחרון של הרכב יpecific.

מקור הנתונים השלישי שלנו הוא מאגר מידע על חברות עסקיות, המכיל פרטים כמו שם החברה, דירוג החברה, הכנסות החברה, שווי החברה ועוד.

ניתן לראות את מאגר החברות העסקית בקטעים בסעיפים [17.6](#) ואת מילון הנתונים [8.5.5.4](#).

אנו משתמשים במידע זה בשבייל להציג את המידע על מקומות העבודה שבהם ציינו הלידים שהם עובדים, ובכך נוכל לחתת לידי עוד מידע נוסף כמו האם החברה שהיא הוא עובד רוחנית, מה שהוא השוק שלו, וכך נוכל לאפיין האם זה משפיע על הסיכוי של הליד להפוך למכירה.

כאשר משתמשים מעלה את קובץ הלידים דרך האתר, קובץ הלידים עולה אל ענן המערכת שלנו שנמצא ב-*Google Cloud Platform* כאשר נוצרת תיקייה ייעודית לכל משתמש לפי השם המשתמש הייחודי שלו.

לאחר שהקובץ הועלה נטענים גם למודל מאגר הנתונים של הרכבים ומאגר הנתונים של החברות, ותהליך כריית המידע מתחילה.

7.1.1.2 תהליכי Pre-processing 7.1.1.2.1 Data Cleaning

בשלב הראשון בחלק ה-Pre-processing נרצה לנוקוט את הנתונים שלנו, חלק זה יבוצע מיד לאחר שהקובץ הועלה למערכת (לאטר).

בהתחלת המערכת בודקת דופליקציות (ערכים כפולים) במפתחות הראשיים של הקובץ שבהם: `lead_id` (מספר זהה ייחודי לכל ליד מקבל בסנייף), ו-`po` (תעודת זהות הלוקוח) אנו נבדוק אם יש ערכים כפולים, ואם יש אנחנו נמחוק אותם.

לאחר מכן ננקה את מאגר הנתונים של החברות, אנחנו נבדוק שאיןערכים ריקים, ונזוז שמות החברות הם בכתב הפה נכונה מסוג ASCII (כלומר מכילים רק אותיות אנגליות ולא אותיות משפטות זרות).

לאחר מכן ננקה את מאגר הנתונים של הרכבים, מכיוון שאנחנו משתמשים במאגר זה בשבייל להצליב מידע על המכוניות שבמהלך התעניינו, אנחנו נמחק ממאגר המכוניות את כל המכוניות שמתבחנת לשנה המינימלית בקובץ הלידים.

אנחנו עושים זאת כי יש לנו מאגר גדול של מכוניות (K343 סוגים שונים של מכוניות) ואני נרצה לחשב את המחיר לכל רכב וגם לצמצם את זמן האלגוריתם. מחיקת המכוניות מתבצעת באופן זמני ולא גורמת למחיקה המידע של כל המכוניות ממאגר הנתונים לצמיות.

7.1.1.2.2 Data Transformation 7.1.1.2.2.1 חישוב מחירי המחיiron של הרכבים

בחלק זהה נמיר ונשנה את סוג העמודות שיתאימו לקלט של האלגוריתם.
בהתחלת נרצה לחשב את מחיר המחיiron של הרכבים בהם התעניינו הלידים, מחיר זה הוא פרמטר חשוב למודל כי על פי אפשר להעריך האם ליד שווה להשקעה או לא.
בשביל לחשב את המחיר ביצענו את הטרנספורמציה הבאה.

יצרנו דатаה פריים סט חדש עליו ביצענו פעולה הקבוצה (`by`) לפי יצרן הרכב, מודל הרכב, ונשנה, לבסוף קיבלנו את הדטהה סט הבא:

Create new df that is group by car_type and year

```
In [104]: df_mean_car_by_model_and_year=df_cars.groupby(['car','model','year']).mean()
df_mean_car_by_model_and_year['price']=df_mean_car_by_model_and_year['price'].map(lambda x: round(x,2))
df_mean_car_by_model_and_year
```

Out[104]:

			id_car	price
car	model	year		
acura	ilx	2013	7.311279e+09	10949.53
		2014	7.313938e+09	14309.54
		2015	7.311765e+09	13216.33
		2016	7.313127e+09	15939.81
		2017	7.312173e+09	15769.86
...
volvo	xc90	2017	7.306397e+09	36431.07
		2018	7.309871e+09	36658.73
		2019	7.314384e+09	49313.17
		2020	7.309088e+09	56328.33
		2021	7.306675e+09	61999.00

3802 rows × 2 columns

כasher האינדקסים הם יצרן הרכב, מודל הרכב, ושם היצור.
העמודות הם `car_id` שהוא ייחודי לכל רכב, ומהירות הממוצע לרכב לפי אותו יצרן, מודל, ונשנה.

בשביל לחשב את מחיר הרכב יצרנו את הפונקציה הבאה:

Calculate the average car price based on the type and the year of the car, if there is no year information, calculate the average based on the type of car

```
In [45]: temp_prices=[]
for index, row in df_leads.iterrows():
    #If the car_type is in df_cars it will enter
    try:
        #Creating a new Dataframe which grouped by the car_type
        new_df=df_mean_car_by_model_and_year.loc[df_mean_car_by_model_and_year.index].loc[row['car_type']].reset_index()
        #If the car_model is in df_cars, will select only the prices of the model
        if row['car_model'] in new_df['model'].values:
            new_df=new_df[new_df['model']==row['car_model']]
            #If there is info of model and year, will get the price of the model and year
            if row['car_year'] in new_df['year'].values:
                temp=new_df[(new_df['model']==row['car_model']) & (new_df['year']==row['car_year'])]['price'].values[0]
                temp=round(temp,2)
                temp_prices.append(temp)
            else:
                #if not, will give the minimum year price.
                temp_prices.append(new_df[new_df['price']==min(new_df['price'])]['price'].values[0])
        else:
            #if there is no info of car_model, will give the car_price by year
            new_df=new_df.groupby(['year']).mean()
            if row['car_year'] in new_df.index:
                temp_prices.append(new_df.loc[row['car_year']]['price'])
            else:
                temp_prices.append(new_df.loc[min(new_df.index)]['price'])

    except:
        #If the car type is not located in df_cars it will calculate price by the car year
        new_df=df_cars.groupby(['year']).mean()
        if row['car_year'] in new_df.index:
            temp_prices.append(new_df.loc[row['car_year']]['price'])
        else:
            #if the year is not located it will calculate by the minimum year
            temp_prices.append(new_df.loc[min(new_df.index)]['price'])

    
```

אנחנו ריצים בולולאה על כל השורות בקובץ הלידים, כאשר בכל איטרציה אנחנו בודקים את יצורן, מודל ונתת הרכב בקובץ הלידים לעומת מאגר המידע של המכוניות שלנו.

אם יצור הרכב נמצא בתוך מאגר הנתונים שלנו, הוא יוצר לנו דатаה סט חדש שבו הוא מסונן לפי אותו יצור רכב, לאחר שהוא יוצר את הדאטא סט המסונן החדש, הוא בודק אם מודל הרכב המבוקש נמצא במאגר, אם מודל הרכב נמצא, הוא בודק האם שנת הרכב הרצויה נמצאת גם כן, אם שנת הרכב לא נמצאת, הוא יחשב את מחיר הרכב לפי ממוצע המחיריהם של היצור ושל המודל.

אם המודול לא נמצא במאגר, אבל יצור ונתת הרכב כן, הוא יחשב את הממוצע על פי היצור והשנה של כל הרכבים במאגר.

אם היצור לא נמצא במאגר, אבל שנת הרכב כן (מקרה זה יכול לקרות אם מסיבה מסוימת הסנייף מחק את הערך של היצור וסימן רק שנה) הפונקציה תחזיר את מחירי הרכב הממוצעים לפי אותה שנת יצור.

בסוף דבר אנחנו נחשב ברמת דיווק גבוהה את מחירי המחרונים של כל הרכבים לפי היצור, מודל רצוי ונתת הרכב.

7.1.1.2.2.2 המרת עמודות התאריך

כאשר ליד נרשם למרכזת, נרשם בקובץ הליד התאריך והשעה שבה הוא הביע התעניניות לגבי הרכב (אם זה בפייסבוק, אינסטגרם או דרך האתר), והוא מציין גם את תקופת ההשכלה שבה הוא מעוניין להסביר את הרכב.

נרצה להמיר את עמודות אלה שיהיו מסוג `datetime`.
נעשה זאת על ידי הפעולה הבאה:

Convert to datetime to proper datatype

```
: df_leads['rental_period']=pd.to_datetime(df_leads['rental_period'],format='%d/%m/%Y')
df_leads['creation_date']=pd.to_datetime(df_leads['creation_date'],format='%d/%m/%Y')
df_leads[['rental_period','creation_date']].head()
```

	rental_period	creation_date
0	2019-06-22	2018-07-21
1	2020-06-07	2018-11-15
2	2020-02-16	2019-05-22
3	2020-01-11	2019-11-08
4	2021-01-26	2019-01-11

7.1.1.2.2.3 חישוב תקופת ההשכרה הרצוייה

לאחר שהמרנו את תקופת ההשכרה, והתאריך שבו נוצר הליד להיות מסוג `datetime`, נוכל לחשב מה כמות הימים הרצוייה שבה הליד מעוניין להשכרה, עכשו ששני העמודות מסוג תאריך נוכל לבצע פעולה חישור פשוטה בשבייל לקבל את הימים.

ניצור עמודה חדשה 'desirable_rental_days' שתכיל את מידע זה.

בנוסף לאחר שהיחסנו את תקופת ההשכרה בימים, צריך לבדוק שלא נוצר מצב שהລוקוח נרשם בתאריך מסוים והוא סימן שהוא רוצה את תקופה ההשכרה לזמן עבר, מה שייגרום למספר הימים הרצויים להיות שלילי, אם המערכת תמצא ערך זהה, אנחנו נחליף את הערכים השליליים במשמעותם הימים הכללי של קובץ הלידים בשבייל לא להשפיע על חישוב האלגוריתם.

7.1.1.2.2.4 החלפת שעת פנית הליד להיות מסוג קטגוריאלי

גילינו שהחלוקת האשכולות מתבצעת בצורה יותר טובה וריך של פונקציית הפוד יורד אם נחליף את שעת פנית הליד למקום ערך מסוים לערך קטגוריאלי.

לכן יצרנו את הפונקציה הבאה:

```
def convert_time_to_categorical(x):
    if (x >= 4) and (x <= 7):
        return 'Early Morning'
    elif (x > 7) and (x <= 11):
        return 'Morning'
    elif (x > 11) and (x <= 15):
        return 'Noon'
    elif (x > 15) and (x <= 19) :
        return 'After Noon'
    elif (x >= 20) and (x <= 23):
        return 'Evening'
    elif (x==24) or (1<=x<=3):
        return "Night"
    else:
        return 'Late Night'
```

כאשר אם הליד פנה אליו בין 4 ל-7 בבוקר, הוא מקבל את הערך 'Early Morning', אם הוא פנה אלינו בין 9 ל-11, הוא מקבל את הערך 'Morning', אם הוא פנה אלינו בין 12 ל-15, הוא מקבל את הערך 'Noon', אם הוא פנה אלינו בין 16 ל-19, הוא מקבל את הערך 'After Noon', אם הוא פנה אלינו בין 20 ל-23, הוא מקבל את הערך 'Evening', אם הוא פנה אלינו בין 00 בלילה ל-3 בבוקר, הוא מקבל את הערך 'Night', ואם הוא פנה אלינו ב-4 בבוקר הוא מקבל את הערך 'Late Night'.

7.1.1.2.2.5 המרת תאריך הלידה לגיל

כאשר ליד נכנס למערכת, הוא מחויב להכנס תאריך לידה בשביל לראות שהוא בגיל המתאים לחתום על חוזה ליסינג נרצה להמיר את תאריך הלידה לגיל מספרי.

Data Merge 7.1.1.2.3

בחלק זה אנחנו יוצרים מיזוג בין 2 הדטה סטים שלנו, דטה סט הלידים ודטה סט החברות המסחריות.

אנו נמנים מבצעים את מיזוג זה כי יש לנו את עמודות הרוח ושווי השוק של החברות בהם עובדים הלידים שיעזרו לנו לסווג האם החברה שבהם הם עובדים היא מבוססת ומכאן יהיה מושלם לשוני הלייסינג לנסוט לסגור חוזה מכירה עם לקוחות זה.

לכן דבר ראשון אנחנו מודאים שעמודות ה-'Market Cap', ו-'profit' הם מסווג כנומיים, אם הם לא אנחנו מmirים אותם להיות מספריים.

לאחר מכן אנחנו יוצרים עותק של דטה סט החברות העסקיות, על מנת לא לבצע שינויים על הדטה סט החברות המקורי.

לפניהם המיזוג, אנחנו נרצה למחוק את כל העמודות שנמצאות בדטה סט החברות שלא רלוונטיות לאלגוריתם ולחישוב, לכן אנחנו נמחק את העמודות הבאות:

`id_company, num. of employees, profitable, rank_change, rank, city, state, newcomer, ceo_founder, ceo_woman, prev_rank, CEO, Website, Ticker, sector, revenue.`

מידע על עמודות אלה ניתן למצאו במילון הנתונים [8.5](#).

לאחר מכן אנחנו נבצע מיזוג, כאשר המפתח עליו נציג יהיה שם החברה.

7.1.1.2.4 Standard Scaler נרמול הנתונים הnumerיים –

נרצה לבצע סטנדרטיזציה נתונים בשביל להתמודד עם ערכים קיצוניים, הסרת הממוצע ושינוי קנה המידה ליחידה אחידה.

צון התקן של ערך מסוים מחושב על ידי הנוסחה הבאה:

$$Z = \frac{X - U}{S}$$

כאשר U זה ממוצע הערבים, ו- S זה סטיית התקן שלהם.

אנו נקבע את נוסחה זו על כל עמודה נומרית בנפרד.

הסיבה שאנו מבצעים זאת על הערבים הnumerיים, כי חלוקת הלידים לאשכולות מבוצעת בצורה יותר טובה כאשר כל הערבים מתפלגים בצורה דומה וממוצע הערבים שווה ל-0.

לאחר שביצענו סטנדרטיזציה נתונים, קובץ הנתונים שלנו מוכן להתקבל כקלט לאלגוריתם.

7.1.1.3 שימוש באלגוריתם

על שימוש באלגוריתם ועל תובנות כריית המידע והמודלים שפותחו מפורט בסעיף [11](#).

7.1.2 חלק שני – אימון האלגוריתמים המונחים (הלקוח מעלה את הקובץ לאחר שאנשי המכירות ביצעו ניסיון מכירה באמצעות הקובץ המקוטלג).

(קישור לקובץ PDF המציג מחברת מסווג Jupyter Notebook המתארת את תהליכי כריית המידע של החלק השני בסעיף [17.6](#)).

בחלק זה, הלקוח העלה את קובץ הלידים בפעם השנייה, לאחר שהוא השתמש בקובץ הלידים המקורי לארבעה קבוצות (לידים "רותחים", לידים "חמים", לידים "בינוניים" לדיים "קרים") סניף הלייסינג ביצע מכירות על קובץ המקוטלג והוא העלה את הקובץ בפעם השנייה כאשר עכשו הוא הוסיף לקובץ עמודה, האם הליד הצליח להפוך למכירה או לא.

בחלק זה אנחנו נתאר את תהליך כריית המידע של הקובץ עם עמודת המכירה, ולבסוף נשתמש באלגוריתמים של למידת מכונה.

7.1.2.1 יבוא הנתונים

כמו בחלק הראשון [7.1.1.1](#) גם כאן השלב הראשון הוא יבוא הנתונים.

אנחנו ניבא את קובץ הלידים החדש לאחר שהועלה למערכת אחריו שנify הלייסינג ביצע מכירה. אנחנו ניבא גם את מאגר החברות המשוחיות כי נרצה בהמשך להשתמש מניפולציות על המידע ולהשתמש במידע נוסף ולהוסיף אותו לאלגוריתם המונחה שיעזר לנו בחיזוי הלידים.

אנחנו מייבאים את המידע מען מערכת Google Cloud Platform של הפרויקט.

7.1.2.2 Pre-processing

נכען תהליך של Pre-processing על קובץ הלידים, מכיוון שחלק מעיבוד המידע כמו חישוב גיל הליד, חישוב מחיר הרכב וכו', בוצע בשלב הראשון, ולכן הרבה משלבים אלו נחסכו בשלב זה, כי המשמש מעלה את הקובץ שהמערכת שלחה ללקוח, לאחר שהוא ביצע עיבוד וטרנספורמציה על המידע.

אך עדין אנחנו צריכים להכין את המידע להתקבל כקלט לאלגוריתם המונחה שכן יש מניפולציות שצירות להיעשות.

Data Cleaning 7.1.2.2.1

המערכת בודקת ערכים כפולים במפתחות הראשיים של הקובץ שהם: lead_id (מספר מזהה ייחודי) שלו ליד מקבל בסנייף), ו-ip (תעודת זהות הלוקוח) אנו בדוק אם יש ערכים כפולים, ואם יש אנחנו נמחק אותם. לאחר מכן, ננקה את מאגר הנתונים של החברות, בדוק שאין ערכים ריקים, ונווידא ששמות החברות הם בכתביה נכון מסוג ASCII (כלומר מכילים רק אותיות אנגליות ולא אותיות משפות זרות).

Data Merge 7.1.2.2.2

כיוון שנרצה לבצע מיזוג של מידע הנמצא במאגר הנתונים של החברות העסקית בין קובץ הלידים, נציג שכבר ביצענו מיזוג מידע ויש בקובץ הלידים המקורי מידע כמו רוח חברתית, ושווי השוק שלו.

אך, נרצה להוסיף גם את הכנסות החברה מכיוון שהמערכת הרatta ביצעים טובים יותר לאחר שהוספנו מידע זה.

7.1.2.3 מחיקת עמודות לא רצויות

לאחר שביצענו את המיזוג, יש הרבה עמודות לא רצויות שਮפריעות למודלים שנפתח, לכן אנחנו נמחק את העמודות הבאות (אך לא לצמיות).

העמודות שנמחק הן:

id, id_lead, first_name, last_name, email, year_of_birth, country, address,
creation_date, rental_period, car_type, company_name, time_catagor, car_model,
segment.

מידע על עמודות אלה ניתן למצוא במילון הנתונים 8.5.5.2.

בסיומו של דבר עמודות כמו תעודת זהות ייחודי, שם משפחה לדוגמה, לא עוזרות למודל לחזות אליו לידים ייפכו למכירה, עמודות אלה ספציפיות מדי לכל ליד ולא מושיפות מידע בעל ערך על שאר המדגם.

7.1.2.4 המרת הנתונים לקטגוריאליים

נרצה להמיר את הנתונים הקטגוריאליים באמצעות שיטת Hot One עליה כתבנו בסקרת הספרות [3.4.2](#).

אנחנו עושים זאת כי האלגוריתמים ממשפחה המונחית כמו רגסיה לוגיסטיות ועכ' החלטה, צריכים לקבל קלט שהוא נומי, ולא מחרוזת, לכן אנחנו נמיר את כל עמודות הקטגוריאליות בשיטה זו.

עמודות לאחר שיטת One-Hot יראו כך:

is_buisness_True	gender_Male	platform_Google	platform_Instagram	platform_Phone	platform_Website	department_Customer service	department_Engineering	d
1	0	0	0	0	1	0	0	0
1	0	1	0	0	0	0	0	0
1	1	0	0	0	1	0	0	0
1	1	0	1	0	0	0	0	0
1	1	0	0	1	0	0	0	0
...
0	0	0	0	0	1	0	1	0
0	1	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0	0
0	1	0	1	0	0	1	0	0

כאשר אם יש את הערך 1 סימן שהיה את הערך זהה בעמודה המקורית, ואם יש את הערך 0, לא היה את הערך זהה בעמודה המקורית.

עכשו האלגוריתמים יוכל לקבל את עמודות אלו כקלט.

7.1.2.5 חלוקת הדטה לדטה סט אימון ודטה סט מבחן

היחס הוא 75% לדטה סט האימון, ו 25% לדטה סט המבחן.

הסיבה שאנחנו עושים זאת היא בשביל שהמודלים של הלמידה המונחית ילמדו את הפרמטרים השונים ובעצם ילמדו מה מאפיין ליד להפיכה למכירה ממשית מדטה סט האימון.

נשתמש בדטה סט המבחן, שהמודל לא למד אותו, בשביל לבדוק את יכולות המודל והאם הוא הצליח לחזות בצורה טובא.

7.1.2.6 שימוש באלגוריתמים של למידת מכונה על הרחבה בשימוש אלגוריתמים ותוכנות כריית המידע מפורט בסעיף 11.

7.1.3 חלק שלישי – המערכת חזה אילו לידים יהפכו למכירה לאחר שנבחר המודל הטוב ביותר (קיים לקובץ PDF המציג מחברת מסוג Jupyter Notebook המתארת את תהליך כריית המידע של החלק השלישי בסעיף 17.6).

לאחר שהמשתמש העלה לנו בשלב השני קובץ שמכיל את עמודת `solds_so`, ככלומר האם הליד הפרק למכירה או לא, כמה מודלים התאמנו על סט הנתונים וחושבו מדדי הדיקוק, כאשר המודל בעל אחוז הדיקוק הטוב ביותר נשמר בענן בתקיית הלוקוח ומוכן לשימוש.

הלוקוח מעלה לנו קובץ חדש המכיל את אותן עמודות כמו הקובץ בשלב הראשון, רק שהקובץ מכיל לידים אחרים, ואנחנו נרצה לחזות אילו לידים יהפכו למכירה.

בשביל זה נדרש לבצע את תהליך ה-Pre-processing שוב על קובץ הלידים החדש, ולהשıp את מחיריו הרכבים.

ורק לאחר מכן, נוכל לחזות באמצעות המודל בעל אחוז הדיקוק הגבוה ביותר אילו לידים יהפכו למכירה.

7.1.3.1 יבוא המודל הטוב ביותר

לאחר החלק השני נרצה להביא את המודל בעל הדיקוק הגבוהה ביותר שיראה אילו לידים יהפכו למכירה. لكن בשלב הראשון ניבא את המודל לסקריפט בשביל שנוכל להשתמש בו.

7.1.3.2 יבוא הנתונים

כמו בחולק הראשון של הפרויקט 7.1.1.1 גם כאן נדרש לייבא את הדטה סטים מהענן.

אנחנו ניבא את שלושת מאגרי הנתונים כי תהליך כריית המידע יצרך להתחילה מנוקודת ההתחלת, מכיוון שהלkoח העלה לנו קובץ חדש לפני שביצענו טרנספורמציה על הדטה סט, נצטרך ליבא את סט הנתונים של החברות ושל הרכבים ולחשב את המידע מחדש.

לכן לבסוף אנחנו ניבא מ-lead סט הלידים החדש.

7.1.3.3 Pre-processing

7.1.3.3.1 Data Cleaning

המערכת בודקת דופליקיציות (ערבים כפולים) במפתחות הראשיים של הקובץ שהם: lead_id (מספר זהה ייחודי שכל ליד מקבל בסוף), ו-ip (תעודת זהות הלוקח) נבדוק אם יש ערבים כפולים, ואם יש נמחק אותם.

לאחר מכן ננקה את מאגר הנתונים של החברות, אנחנו נבדוק שאין ערבים ריקים, וכן אדעת שמנות החברות הם בכתביה נכון מסוג ASCII (כלומר מכילים רק אותיות אנגליות ולא אותיות משפטות זרות).

לאחר מכן ננקה את מאגר הנתונים של הרכבים, מכיוון שאנחנו משתמשים במאגר זה בשבייל להצליב מידע על המכוניות שביהם הלידים התעניינו, אנחנו נמחק ממאגר המכוניות את כל המכוניות שמתוחת לשנה המינימלית בקובץ הלידים.

7.1.3.3.2 Data Transformation

7.1.3.3.2.1 חישוב מחירי המכירות של המכוניות

نبצע את חישוב המחיר כמו שעשינו בחלק הראשון אשר מפורט כאן [7.1.1.2.2.1](#).

7.1.3.3.2.2 הרמת עמודות תאריך

نبצע את המרות סוג עמודות להיות מסוג datetime כפי שפורסם כאן [7.1.1.2.2.2](#).

7.1.3.3.2.3 חישוב תקופת ההשכרה הרצiosa

נחשב את תקופת ההשכרה הרצiosa כפי שפורסם כאן [7.1.1.2.2.3](#).

7.1.3.3.2.4 הרמת תאריך הלידה לגיל

נחשב את הגיל לפי תאריך הלידה כפי שפורסם כאן [7.1.1.2.2.5](#).

7.1.3.4 Data Merge

نبצע את אותו מיזוג בין דטה סט הלידים החדש לבין מאגר הנתונים של החברות העסקיות על מנת לקבל עוד מידע שימושי להשתמש בו כקלט לאלגוריתם, כפי שפורסם כאן [7.1.1.2.3](#).

7.1.3.5 הרמת הנתונים הקטגוריאליים

נרצה להמיר את כל העמודות הקטגוריאליות באמצעות שיטת Hot-One כדי שיוכלו להתקבל אצל המודל למידה הטוב ביותר שנבחר בשלב השני כפי שפורסם כאן [7.1.2.4](#).

לאחר שלב זה הדטה סט שלנו סיים את תהליך Pre-processing ויהי אפשר לבצע עליו תחזיות

7.1.3.6 שימוש במודל המונחה הטוב ביותר ותוצאותיו

על הרחבה בשימוש אלגוריתמים ותוצאות כריית המידע נפרט בסעיף 11.

7.2 אפיון המערכת – מערכות מידע

7.2.1 טבלת בעלי עניין

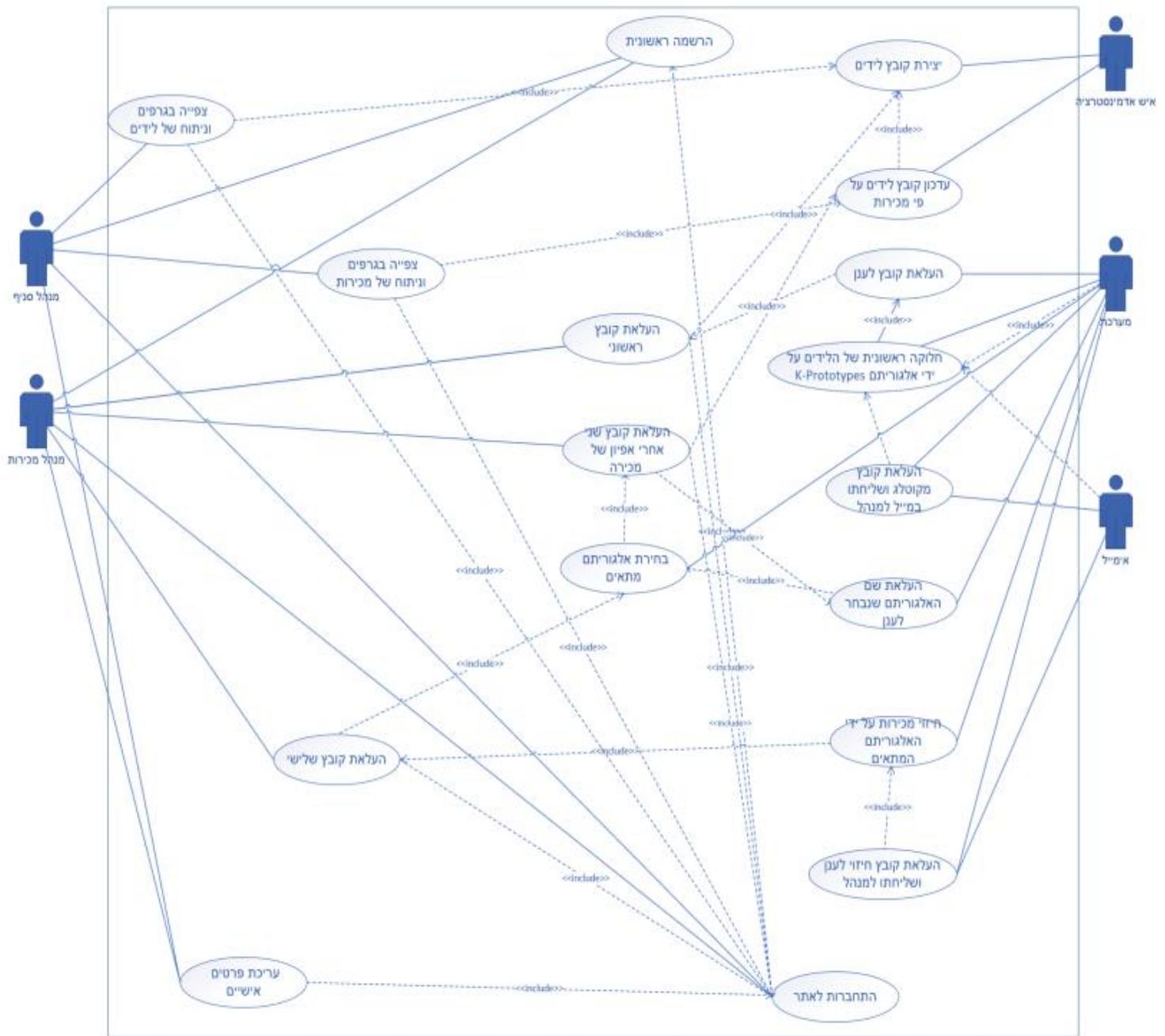
בעלי עניין	ספקים	הגורם מהפתרון המוצע	התיחסות	יכולות ומאפיינים	אילוצים וمبرליות
-	המערכת עתידה להגדיל את המכירות של החברה, משמע, יותר רכישות ממהספקי השוניים	תמייה, הפתרון מציע הגדלת רווחים גם הספקים מרווחים מהם	אין מעורבות בתהליכי המערכת	-	-
-	חברות מתחרות	הפסד של מכירות פוטנציאליות, המערכת עשויה לשפר את חווית הלקוח	התנגדות	אין מעורבות בתהליכי המערכת	-
-	בעלי החברה	הגדלת הכנסתות בזכות המערכת	תמייה, הדרג הגבווה בחברה מעוניין להגדיל את הכנסתות החברה	תשולם על השירות של המערכת	-
חברות ביטוח תלויות בחברת הליסינג	חברות ביטוח עשויה להגדיל את רווחי המכירות, שבחורות הביטוח גם יכולות במחיר של המכירה	חברות הביטוח יכולים לסמוך בהם בחלוקת הון גוףם גם לאירועים נזקים, המשמשים כמקור להרחבת השירותים	מכירת הקובץ עברו לחברות הביטוח ומהוות תלויות בחברת הליסינג	חברות הביטוח יוכלו לשמש בקובץ הילדים הממוני ולמכור להם את פוליסות הביטוח גם ללקוחות שלא רכשו את הביטוח רכישת הליסינג	-
בעלי החברה	הגדלת הגדלת המכירות בזכות הגדלת הבונוסים החודשיים בזכות המערכת	יעודד שימוש במכוון המלצה שמקבל ממערכת המלצה	לא יכול לשמש במערכת אלא רק בתורים שמתאפשרים במרקם מסמך"ל המכירות	כמפורט בסעיפים	-
ליך מהיר ואישי יותר	קיבלה שירות באופן מהיר ואישי יותר	תמייה, יגער לבקשתו מהר יותר	אין מעורבות בתהליכי המערכת, התהילה רק שkopf מבחינת הלקוח	כמפורט בסעיפים מהיון ישנים לקוחות שיקבלו תיעוד אחרון למורות פניהם המוקדמת	-

טבלה 1.7.2.1. בעלי עניין

הגורם	תועלת מהפתרון המוצע	ערך עיקרי /	יכולות ומאפיינים	התיכוןות	אלוצים ומגבליות
מנהל מכירות	עליה קובץ	תמייה ושימוש נרחב במערכת	קביעת פרמטרים למערכת וניהול הנתונים במערכת	-	-
מנהל סניף	הבנה של מצב המכירות ונתוני הלקוחות והילדים	תמייה בקבלת תמונה רחבה במערכת	צפיה בחלוקת הילדים והמכירות על ידי הגրפים שיש של המצב	לא יכול לעורוך את הנתונים במערכת אלא רק לצפות בדוחות	שחקן
איש אדמיניסטרציה	חולץ זמן בחלוקת הילדים לאנשי המכירות	שולח את הילדים שאוסף למנהל המכירות	אין לו יכולות פיזיות במערכת, הוא אוסף את הילדים ושולח אותם לאיש	אין לו גישה ישירה למערכת, הוא באינטרנט אקדמי אייתה רק מכיוון שהוא צריך להעביר את קובץ הילדים למנהל המכירות.	

טבלה 7.2.2. שחקנים

7.2.3. תרשים מקרי שימוש – Use Case



תרשים 7.2.3. מקרי שימוש

7.2.4 טבלת מקרי שימוש

מו"ד	מקרה שימוש	מטרת התהילה	שחקן ראשי	תיאור קצר
1	הרשמה ראשונית	שמירת פרטי המנהל כולל פרטיים אישיים ופרטי התchapרות	מנהל סניף ומנהל מכירות	המנהל מזין פרטיים אישיים, פרטי השירות, פרטי הסנייפ ופרטי התchapרות DB והפרטים נשמרים בatabase "יעודי"
2	התchapרות לאתר	התchapרות למשתמש קיים וצפיה בנתוני העבר. לא התchapרות לא ניתן להשתמש במערכת	מנהל סניף ומנהל מכירות	הΖנת שם משתמש וסיסמה של אותו המנהל
3	העלאת קובץ ראשי	סיווג ראשוני של קובץ הלידים שיש בידיו המנהל	מנהל מכירות	המנהל מעלה את הקובץ לידיים הראשוני בפורמט CSV על ידי לחיצת כפתור מהלך
4	העלאת קובץ לען	שמירה של הקובץ בענן "יעודי"	מערכת	המערכת מתממשקת לAPI של גוגל ומעלה את הקובץ שהתקבלה מהלך
5	חלוקת ראשונית של הלידים על ידי אלגוריתם - K Prototypes	חלוקת לקבוצות של ידיים שהוגדרו מראש על ידי האלגוריתם	מערכת	המערכת מזהה את קובץ הלידים שהועלה ומתחלילה בהפעלת אלגוריתם "יעודי" אשר משמש כ- Unsupervised Learning
6	העלאת קובץ לען ושליחתו למנהל	סיום התהילה הראשוני ועדכן המנהל על ידי מייל ושמירת הקובץ בענן לשימוש עתידי	מערכת ואימייל	המערכת מעלה את הקובץ לען גוגל ונשלח אימייל שמצויה עם שם המזוכר על ידי API של GMAIL
7	העלאת קובץ שני אחרי אפיון של מכירה	עדכן האלגוריתם אם בוצעה מכירה	מנהל מכירות	המנהל מעלה בלחיצת כפתור קובץ CSV בעל עמודות בוליאנית המתארת אם הליד הפרק למכירה או לא
	בחירה אלגוריתם מתאים	בחירה האלגוריתם המתאים ביותר עבור אותו קובץ לידיים	מערכת	המערכת משתמש במספר אלגוריתמים של מידע מונחיות ובוסף בוחרת את האלגוריתם עם אחוז הדיווק הגבוהים ביותר

המערכת מתממשקת לAPI של גוגל ומעלה את השם לענן "יעוד" של המנהל	מערכת	על מנת לשמור את פרטיו האלגוריתם שנבחר, המערכת תעלה את השם לענן	העלאת שם האלגוריתם שנבחר לענן	8
העלאת קובץ לידיים בפורמט CSV ללא עמודות נמכר או לא נמכר על ידי לחיצת כפתור	מנהל מכירות	להלן מעלה קובץ לידיים ללא תוצאות על מנת שהמערכת תמיין לו אוטומטית בדיקת הגבואה ביותר על ידי האלגוריתם שנבחר	העלאת קובץ שלישי	9
המערכת מפעילה את האלגוריתם שנבחר	מערכת	הפעלת האלגוריתם	חיזוי מכירות על ידי האלגוריתם המתאים	10
המערכת מתממשקת לAPI של גוגל ומעלה את הקובץ לענן תיוקיה "יעודית" של הלקוח ושולחת מייל על ידי GMAIL הכלול בתוכו את הקובץ המופיע	מערכת ואימייל	סיום התהליך של הלמידה המונחית ושליחת הקובץ למנהלו. בנוסף, שמירה בענן לצורך תיעוד ושימוש עתידי	העלאת קובץ מוקטן לענן ושליחתו למנהלו	11
עדכן פרטיים אישיים או פרטי התחברות	מנהל מכירות ומנהל סניף	שמירה על רלוונטיות ועדכניות- במידה והמנהל רוצה הוא יכול לעדכן פרטיים אישיים בכל רגע נתון	עריכת פרטיים אישיים	12
איסוף פרטיים לידיים מפלטפורמות שונות: פיסבוק, אינסטגרם, אחר אינטראקט "יעוד" ... ושמירה בקובץ CSV המקורי	איש אדמיניסטרציה	יצירת קובץ הלידים מפלטפורמות שונות לקובץ אחד מאוחד לצורך הטענתו למערכת	יצירת קובץ לידיים	13
קיבלה מידע מהאנשים המכירות אם הליד הפר למכירה או לא ועדכן בקובץ לידיים בפורמט CSV המקורי	איש אדמיניסטרציה	עדכן אותו קובץ לידיים ראשוני אם הליד הפרק למכירה או לא לצורך שימוש בו בחילק של הלמידה המונחית	עדכן קובץ לידיים על פי מכירות	14
העלאת קובץ לידיים לפני שידוע אם נמכר או לא ולחייב על גרפ נבחר	מנהל סניף	ניתוח פרטי לידיים של הסניף תוך צפיה בגרפים "יעודים" עבור אותו הסניף	צפיה בגרפים וניתוח של לידיים	15
העלאת קובץ לידיים אחרי התוצאות ולחייב על גרפ נבחר	מנהל סניף	ניתוח פרטי לידיים והתוואה שלהם על ידי גרפים "יעודים"	צפיה בגרפים וניתוח של מכירות	16

7.2.5 דרישות פונקציונליות (רשימה ממושפרת במבנה היררכי, לפי נושאים)**7.2.5.1 דרישות מידע**

7.2.5.1.1 המערכתبعثת

- העלאת קובץ לדיים נשמר במסד הנתונים את השעה שבה הועלה הקובץ, שם הקובץ ושם המשמש.

7.2.5.1.2 אחסון קובץ הלידים בענן בתיקייה ייעודית של המשמש.

7.2.5.1.3 שמירת פרטי המשמש והסניף בטבלאות במסד הנתונים בעת הרשמה למערכת.

7.2.5.1.4 סוג הנתונים המתקבלים למערכת, יהיו מותאמים לדרישות המערכת שהוגדרו מראש (*).

7.2.5.1.5 הטבלאות במסד הנתונים יוגדרו כדינמיות ויהי ניתן לשנות את הנתונים כתלות בעדכוני המשמש למערכת.

7.2.5.1.6 אפיון עובדים וסניף - Nice to have.

7.2.5.2 דרישות פעולה

7.2.5.2.1 המערכת תאפשר לקבל קבצי CSV ב3 התהיליכים של העלאת הקבצים. (בדרישות ההתחלתיות הגדרנו קובץ אחד בלבד ושלב אחד).

7.2.5.2.2 המערכת תזהה אם הקובץ שהועלה בעמוד ה-*Dashboard* מכיל בתוכו את העמודה של נמוך או לא ולפי זה תציג עמוד גרפים מותאים.

7.2.5.2.3 *Dashboard* יציג גרפים המנתחים את הלידים ואת המכירות. (בדרישות ההתחלתיות התייחסנו לניטוח רק של המכירות- אך ראיינו שניטוח של הלידים יוכל לעזור גם הוא להסקת מסקנות).

7.2.5.2.4 המערכת תבצע אלגוריתם ממפחית *Unsupervised Learning* על הקובץ הראשוני שהועלה למערכת.

7.2.5.2.5 המערכת תבצע אלגוריתם ממפחית *Supervised Learning* על הקובץ השימושי שהועלה למערכת. (בדרישות המקוריות כתבנו שהיה זה הקובץ השני שעליו תבצע מידיה מונחית).

7.2.5.2.6 המערכת תעלה לען על ידי התממשקות ל*API* של גוגל את הקבצים שהמשמש העלה. -הוסף דרישת.

7.2.5.2.7 המערכת תשלח מייל למשמש עם הקובץ לאחר ביצוע הלמידה המונחית והלמידה הבלתי מונחית עם הקובץ המקורי. -הוסף דרישת.

7.2.5.2.8 המערכת תשלח קבצים עם תוצאות וסבירים של הקטגוריות לאחר הפעלת האלגוריתם של הלמידה המונחית. - הוסף דרישת.

7.2.5.3 דרישות ממשך

7.2.5.3.1 המערכת תחייב הרשמה ראשונית באופן ידני למשמש הכלל פרטיים אישיים ופרטית החברות ופרטית סניף.

7.2.5.3.2 המערכת תאפשר החברות רק למשמשים שנרשמו.

7.2.5.3.3 המערכת תבקש הגדרה של שם משתמש ייחודי וסיסמה עבור המשמש החדש.

7.2.5.3.4 המערכת תאפשר להגדיר את המבנה הארגוני של הסניף ותבקש שיווק של אנשי המכירות המשווים לסניף תוך התייחסות ל-*Relationships* של הישויות. - Nice to have.

7.2.5.3.5 המערכת תדרש העלאת קובץ ראשי של ידיים לביצוע ניתוח ראשוני.

7.2.5.3.6 המערכת תציג גרפים על פי הקבצים שהועלו ל מערכת *Dashboard* המיועד למנהל המכירות לניתוח של הלידים והמכירות.

- 7.2.5.3.7 המשתמש יוכל לבצע ערכות על פרטי האישים ופרטיו התחברות שלו.
- 7.2.6 דרישות לא פונקציונליות
- 7.2.6.1 דרישות ביצועים
- 7.2.6.1.1 זמן הצגת התוכן הראשון לא יעלה על 3 שניות.
- 7.2.6.1.2 גיבוי בסיס הנתונים יבוצע בתדרות יומית.
- 7.2.6.1.3 קובץ הלידים של כל לקוח לא יעלה על MB10.
- 7.2.6.1.4 המערכת תהיה נגישה לשימוש רק ב-Web.
- 7.2.6.1.5 המערכת תהיה נגישה לכל מערכת עם גישה לאינטרנט וכונן זיכרון.
- 7.2.6.1.6 המערכת תאפשר העלאת קובץ אחד בכל שימוש (ולא העלאת כמה קבצים ביחד).
- 7.2.6.1.7 המערכת לא תהיה מוגדרת מייד למשתמש, אלא רק כדי למיון קבצים קיימים והציג הנתונים מהקבצים הללו.
- 7.2.6.2 דרישות אינטראקטיביות
- 7.2.6.2.1 המערכת תאפשר עריכת הרשאות למשתמשים באופן ממוקד.
- 7.2.6.2.2 הרשאה למנהל המערכת.
- 7.2.6.2.3 הרשאה לדרג ניהול
- 7.2.6.2.4 עריכת הרשאה מותאמת אישית.
- 7.2.6.2.5 המערכת תהיה זמינה בכל שעות היוםה.
- 7.2.6.3 דרישות אבטחת מידע
- 7.2.6.3.1 המערכת תבקש מהמשתמש סיסמה ייחודית בעלת 8 תווים לפחות הכוללת מספרים ואותיות.
- 7.2.6.3.2 בעת יצירת הסיסמה הייחודית, המשתמש יתבקש לאמת את הסיסמה.
- 7.2.6.3.3 הסיסמה תהיה מוגדרת ב-HASH במערכת במידה ותהיה בעית אבטחה, כך לא יחשפו נתונים המשתמשים.
- 7.2.6.3.4 רישום של המשתמשים במערכת וזמן פעילותם - Nice to have.
- 7.2.6.3.5 תיעוד פרטי המשתמש.
- 7.2.6.3.6 תיעוד זמן שימוש - Nice to have.
- 7.2.6.3.7 תיעוד פעולות המשתמש - Nice to have.
- 7.2.6.4 דרישות חומרה
- 7.2.6.4.1 מערכת המידע תאופין בגמישות בהתאם לדפדים השונים.
- 7.2.6.4.2 מערכת המידע דורשת שהיא יכולה דוח לידיים המכיל לפחות 500 שורות של לידיים.

7.3 ניתוח חלופות טכנולוגיות

7.3.1 הצגה של מספר חלופות מערכתיות

בשער זה אנו נציג את כל האלטרנטיבות להקנת המערכת שלנו מבחינה טכנולוגית, ישנים המונע דרכים לIMPLEMENTATION המערכת על בסיס אחר שיקבל נתונים ויחזיר נתונים למשתמש. באמצעות סקר הספרות והמחקר שערכנו ונתקדם בכמה טכנולוגיות מרכזיות ובאמצעות טבלה שמראה את הקритריונים החשובים לכל טכנולוגיה נגיע להחלטה לגבי הפלטפורמות בהם נשימוש. בטבלאות לבחירת החלופות השונות ניתן דירוג בין 1 ל-10. כאשר 1 - אינה עונה כלל על קритריון זה, ו-10 – עונה בצורה מצינית.

7.3.1.1 חלופות לשרת

kritirion	משקל	הסבר	שרות בענן	שרות פיזי
עלות	25%	אנו נ שאף לעליות נמוכות בשבייל להוזיל את הפROYIKET	8	5
גמישות	15%	השרת שבו נבחר צריך להיות גמיש לשינויים, במידה וכמות המידע גדלה נוכל בקלות להגדיל את מקום האחסון	9	2
התממשקות	20%	שהשרת יוכל להתממשק בקלות עם פלטפורמות שונות	9	6
אבטחת מידע	15%	רמת אבטחת המידע של השרת	5	9
מקום אחסון (פיזי)	10%	מקום פיזי של השרת עשוי להיות קשיים (קירור, חדר שירותים, וכו').	10	4
ניהול סיכונים	15%	תלות בגין חיצוני שיכול להשפיע על תפקוד השרת (קריסת השרת, תקלות)	5	6
סה"כ	100%	-	7.65	5.4

7.3.1.1.1 חלופות לשרת

החלופה הנבחרת היא שרת בענן, באופן כללי רוב החברות המובילות בשוק מעבירות את מסד הנתונים שלהם לענן מכיוון שהוא מהיר יותר וזול יותר.

גם אנחנו בפרויקט שלנו נבחר להשתמש בענן שם נ אחסון את הקבצים שיועלו על ידי המשתמש, בשל מגבלת התקציב של הפרויקט נשימוש ב-Google Cloud Platform שהוא קל וnoch ופשוט לשימוש באמצעות ספריות יי'ודיות ופונקציות API.

בנוסף, קבצי ה-CSV שלנו לא כבדים מדי, לכן לא צריך הרבה מקום לאחסון וכן שימוש בענן יהיה זול יותר, הרו שימוש בענן הוא על בסיס שימוש בפועל, ושרת פיזי הוא מקום אחסון שנקבע מראש ולא דינמי לצורך של הארגון.

7.3.1.2 פלטפורמת לבניית אתר

קריטריון	משקל	Django	Flask	Wix
עלות	10%	10	10	4
נוחות שימוש	25%	7	8	10
נדרש השלמת פערים	20%	6	8	9
זמן ומורכבות פיתוח	20%	3	7	8
קהילה מפתחים	10%	9	9	5
ציהוי וapurion שגיאות בritchah	15%	6	8	3
סה"כ	100%	6.35	8.1	7.25

7.3.1.2. פלטפורמה לבניית אתר

הפלטפורמה שבה בחרנו להשתמש היא Flask שהיא ספרייה מובנית בפייתון לפיתוח אתרים.

באופן כללי התלבטנו היכן נפתח את האתר שלנו ועם באיזה כלី להשתמש, סופג Django הוא גם ספריית פיתון לבניית אתרים, לעומת זאת Wix הוא פלטפורמה ידועה ונוחה לבניית אתר בקלות ללא צורך ידע טכניות.

מבחןת עלות שתי הספריות קיבלו ציון גבוהה מכיוון שהן חינמיות לשימוש לעומת Wix שעולה כסוף ואנחנו שואפים לצמצם כמה שיותר את הוצאות הפ羅יקט.

יש בשתי הספריות צורך להשלמת פערים, באופן כללי פלטפורמת Flask לפיתוח האתר היא הרבה יותר פשוטה ונוחה לעומת Django, כאשר אתה רוצה שהאתר שלך יהיה "עדי" Scalability" עדיף להשתמש ב-Django אך מכיוון שאנחנו יודעים שהאתר שלנו לא צריך לכלול פונקציונליות גבוהה בחרנו להשתמש ב-Flask.

בקטגורית השלמת פערים Wix קיבלת את הציון הכי גבוהה מכיוון שהיעוד שלו הוא לבנות אתר ב מהירות וקלות.

מבחןת זמן ומורכבות פיתוח כמו שציינתי בסעיף הקודם Flask קלה ונוחה לשימוש הרבה יותר לעומת Django, Wix כמובן יותר קלה אך מוגבלת מבחינת הרצת קוד ואלגוריתמים בהם נרצה להשתמש.

לסיכום, מכיוון שאנו נפתח את האלגוריתמים שלנו בפייתון בחרנו להשתמש בפלטפורמת פיתוח פיתון, כאשר יהיה יותר נוח ופשוט לפתח הכל על אותה פלטפורמה, השלמת פערים ונוחות שימוש הינו פקטורי מרכז בהחלטה שלנו ולמרות שבקטגוריות אלה Wix קיבלת ציון יותר גבוה מ-2 האלטרנטיבות, יש לו חסרונות בдинמיות ולכן לא בחרנו בו אלא ב-Flask.

7.3.1.3 מסד נתונים לאתר

החלופה הנבחרת למסד נתונים לאתר היא SQLite. הקriterיוונים ששמנו בעדיפות עלינה היא העלות ויכולת התאמושקות של מסד הנתונים עם שפת התוכנה של האלגוריתם. בנוסף, אחסון מסד הנתונים ב-SQLite הרבה יותר דינמי וgemäß לשינויים היות והמבנה שמור בקבצי האפליקציה אשר מKENה גם אבטחת מידע גבוהה יותר היות ולא ניתן לגשת לפרטיה האפליקציה ללא התקנות נוספות.

Oracle		MySQL	SQLite	משקל הקריטריון	קריטריונים
3	6	10	30%	עלות	
6	7	10	20%	התקנה	
6	8	10	30%	גמישות לשינויים ודינמיות	
6	7	9	20%	התאמה לקוד	
5.1	7	9.8	100%	סה"כ	

טבלה 7.3.1.3 מסד נתונים לאתר

7.3.1.4 פלטפורמת אחסון ענן

למרותSCP של החלופות נוחות, החלופה הנבחרת למסד נתונים לאתר היא Google Cloud Storage ובניגוד ל-Google Drive אפשר שימוש יותר עסקי מאשר אישי. בנוסף, גוגל מאפשר להשתמש בפונקציות API המתאמושקות בקלות עם Python ויצרים סביבה נוחה עבור מתקנתים וחברות קטנות.

One Drive		Google Cloud Storage	Google Drive	משקל הקריטריון	קריטריונים
8	8	8	8	30%	עלות
9	10	9	5%	התקנה	
6	10	7	30%	קלות התאמושקות	
9	10	9	20%	גודל אחסון	
8	10	8	15%	למידה עצמי	
7.65	9.4	7.95	100%	סה"כ	

טבלה 7.3.1.4. פלטפורמת אחסון ענן

7.3.1.5 שפת פיתוח

אנחנו לא נשימוש בטבלת קriterיוונים כי שפת הפיתוח היחידה שלקחנו בחשבון היא פיתון. פיתון היא שפת פיתוח דינמית פשוטה לשימוש לעומת אחרות ויש בה קהילת מפתחים גדולה. היתרון הגadol שלו שהוא שימושית גם בעולם-h Data Science שבו עוסק הפרויקט שלנו, כלומר יש לה הרבה ספריות ייודיות שניתן ליבא בהם אלגוריתמים ולעשות בהם שימוש. בנוסף גם שפת פיתוח-h Web שלא היא רחבה ובעצם אלו שתי הפעולות העיקריות שאנו צריכים להשתמש בהם בפרויקט.

בחינה הפרקטית, למדנו במהלך התואר קורסים בפייתון וкриיט מידע על בסיס פיתון, لكن זו הייתה בחירה אידיאלית ונוחה על מנת ליישם זאת בפרויקט

7.3.2 הצגת חלופות השונות לשימוש באלגוריתם

עיקר הפרויקט הינו שימוש באלגוריתם אשר יחזא את הלידים שיהפכו למכירה ממשית וכן אב הטיפוס יכול אלגוריתם משפחתי הלמידה המונחית ומהלמידה הבלתי מונחית. לפיכך, נרצה לבצע את השוואת בין האלגוריתמים השונים לטובות בחירות המודל הטוב ביותר.

לצורך כך נבחן מספר אלגוריתמים משפחתי הלמידה הבלתי מונחית.

7.3.2.1 חלופות לאלגוריתמים משפחתי Learning-h-unsupervised

שקלנו בעבודה לבחור בכמה אלגוריתמים ולבסוף נשארנו עם כמה בחירות אופציונליות:

K-Prototypes Clustering	Fuzzy K-means	K-Means Clustering	משקל הクリיטריון	קריטריונים
7	6	8	30%	יעילות
8	6	8	30%	נדרש השלמת פערים
10	10	8	20%	זמן ומורכבות פיתוח
10	7	8	10%	תוצרת קבלת הנתונים
9	9	9	10%	קהילה מפתחים ומחוקרים
8.4	7.2	8.1	100%	סה"כ

טבלה 7.3.2.1. חלופות לאלגוריתמים הבלתי מונחית כשםדובר על אלגוריתמים מלמידה בלתי מונחית, היו כמה אופציות רלוונטיות אך בסופה של דבר רצינו ליצור מיוון ראשוני לדאותה לא מתואג בשביל להביא אותו לאנשי המכירות על בסיס מאפיינים דומים.

הרחבנו על אלגוריתם Fuzzy Algorithm [3.5.4](#) בסקירת הספורות K-mean clustering הוא גם משפחת האלגוריתמים של "Clustering" ולפי השם ניתן להבין שהמהות שלו דומה והוא מאוד דומה לאלגוריתם K-means, ההבדל היחיד הוא שבמקרה להקצות נקודה בלבד לאשכול אחד בלבד, הנקוצה יכולה להיות "מטושטשת" ככלمر נקודה יכולה להיות שייכת בין שני אשכולות או יותר.

Fuzzy K-Means Algorithm, K-Means Algorithm, מוגדר כאשכול קשה, שבו כל אחת מהנקודות שייכות לאשכול אחד, מחליף את האשכולות הרכבים יותר לחיפוי.

נקודה בודדת באשכול רק יכולה להיות שייכת ליותר מאשר אשכול אחד עם ערך זיקה מסוים כלפי כל אחת מהנקודות.

הזיקה היא ביחס למרחק של אותה נקודה ממרכז האשכול. בדומה ל-K-means, Fuzzy K-means עובד על האובייקטים שמיידת המרחק מוגדרת להם וניתן לייצג אותם במרחב הווקטור ה- n -ממדי.

לאחר שכתבנו והסבירנו על K-Means Clustering, K-Fuzzy Algorithm, ועל K-Prototypes Algorithm בפוסט להשתמש ב-

הבעיה המרכזית של K-Means Clustering-ו-K-Fuzzy Algorithm, שהם לא יכולים לקבל קולט איקסים אשר מכילים עמודות לא נומריות.

לכן היתרון הגדול של K-Prototypes שהוא משלב במרקח ובפונקציית ההפסד את מරחק האי דמיון לעמודות הקטגוריאליות, והמרקח האוקלידי לעמודות הnumerיות.

לכן לשיכום, מכיוון שהدادטה שלנו מכיל עמודות נומריות וקטגוריאליות, ולא רצינו לפגוע בבדיקה האלגוריתם ולהפוך את העמודות הקטגוריאליות לנומריות או להפוך, החלטנו להשתמש באלגוריתם K-Prototypes אשר נתן לנו תוצאות אינטואיטיביות וחילק לנו את הלידים בצורה טובה.

7.3.2.2 חלופות לאלגוריתמים ממשפחה ה-Supervised Learning

Random Forest	Decision Tree	Logistic Regression	משקל הקритריון	קריטריונים
9	8	9	30%	יעילות
8	8	8	30%	נדרש השלמת פערים
10	7	8	20%	זמן ומורכבות פיתוח
6	9	7	5%	תצורת קבלת הנתונים
10	10	10	15%	קהילה מפתחים ומחוקרים
8.9	8.15	8.55	100%	סה"כ

7.3.2.2 אלגוריתמים ממשפחה ה-Supervised Learning

באופן כללי התלבטו על מספר אלגוריתמים שונים גם כי בתחום הלמידה המונחית יש המון אופציית לסייע Attributes, וגם כי נרצה לבחור את האלגוריתם בעל תוצאה הדיק הגדולה ביותר. הפרמטרים שלקחנו בחשבון בעת בחירת האלגוריתם:

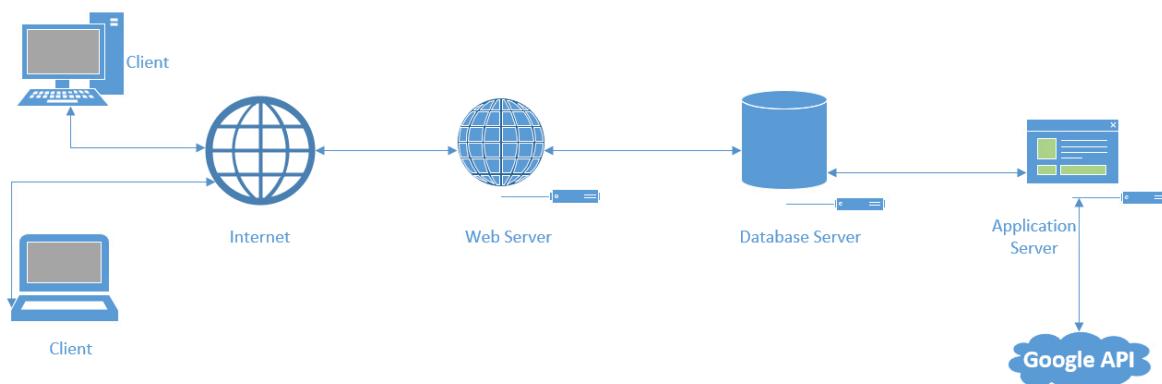
- **יעילות** – כל האלגוריתמים שבחרנו ייעילים הן מבחינת זמן ריצה והן מבחינת הטמעה במערכת.
- אנחנו משתמשים במודלים מוכנים אשר באמצעות API אנחנו מיבאים את הפונקציה המתאימה ומכוונים את הפרמטרים המתאימים.
- **השלמת פערים** – על כל האלגוריתמים האלה למדנו במכילה בקורס כריית מידע, אך גרסיה לוגיסטיבית, עץ החלטה ו"עיר רנדומלי" הם יותר פשוטים להטמעה וגם כאן נושא הסיבות הוא פרמטר חשוב, لكن בחרנו לשימוש באלגוריתמים אלה.
- **זמן ומורכבות פיתוח** – גרסיה לוגיסטיבית ועץ החלטה הם יחסית פשוטים להטמעה, גם Random Forest, הבעה אותו שהוא מօסיף סיבוכיות למודל כי הוא יוצר מאות עצים שונים, אבל מכיוון שפוחיקט שלו מספר הleafים הוא לא גדול, החישוב מתבצע במהירות כמו שאר המודלים.
- **תצורת קבלת הנתונים** – עץ החלטה הוא אלגוריתם מצוין להראות איך בכל צומת התקבלה ההחלטה להמשיך לצומת הבא, וניתן אחר כך להראות זאת בצורה ויזואלית, בגרסה לוגיסטיבית ניתן לקבל את ההסתברות של הליד להיות שיר לקבוצה מסוימת. בוגר ל- Random Forest, מכיוון שהוא יוצר המון עצים זה מסביר את הויזואליות ולכן הוא קיבל ציון נמוך בפרמטר זה.
- **קהילה מפתחים ומחוקרים** – על כל האלגוריתמים יש המון מידע באינטרנט لكن נתנו לכולם ציון 10 כי היום קיימים דוקומנטציה על כל אופן החישוב והשימוש בקוד עצמו באמצעות ספריות ייעודיות.

לסיכום: לאחר שסקרנו את כל הפרמטרים ראיינו לפ' התוצאות כי Random Forest הביא את תוצאות הדיקט הגבואה ביותר, אך, היית ומדובר במערכת המלצה ולמידת מכונה, ברגע שנכנייס דатаה חדש, תוצאה הדיקט יכולה להשתנות ומודל אחר יכול להתאים לאותו דטה סט.

ולכן, בחלק של ה- Supervised Learning האלגוריתם יקח את המודל בעל תוצאות הדיקט המשוקללות הטובות ביותר.

8 תיכון המריכת – System Design

Network Diagram 8.1

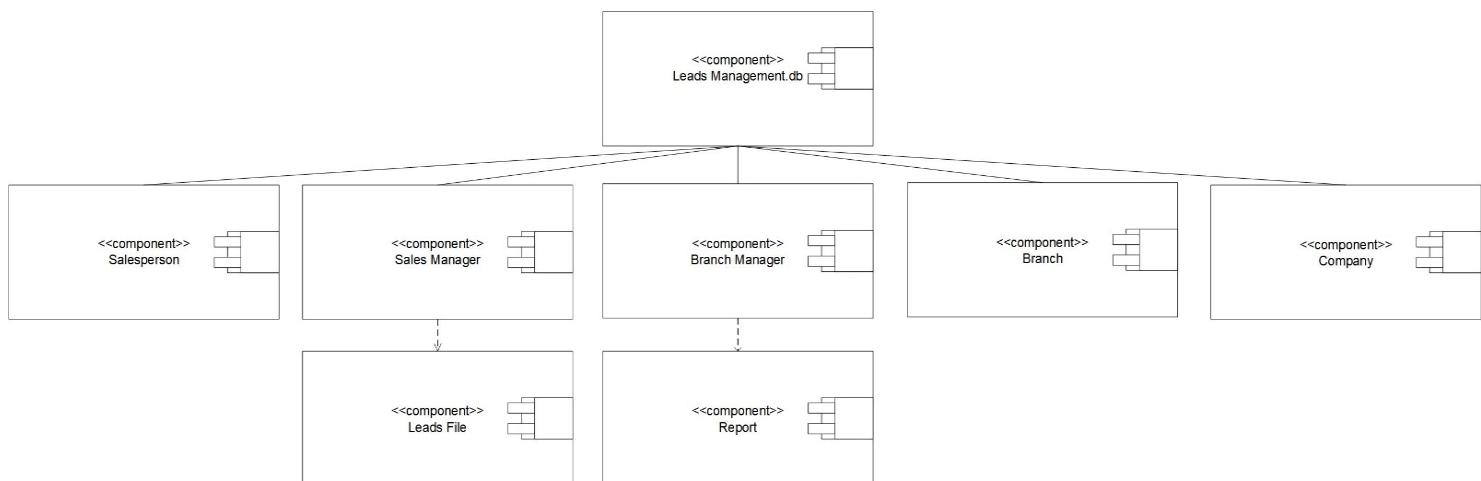


תרשים 8.1 Network Diagram

- I. Client - שכבה הדפדף מהוות שכבה תצוגה למשתמש אליה המשתמש יכנס למערכת ויכול לבצע את הפעולות השונות. אנו מצפים שהמשתמש ישמש במחשב של העובודה שלו לצורך העלאת קבצי הלידים וצפיה בגרפים המתאימים ולכך המערכת תהיה מותאמת לWEB בלבד בכך שלא יעשה שימוש אישי מחוץ לשעות העבודה דרך הסלולר. המנהל משתמש בRouter או ב-WIFI על מנת להתחבר לאתר.
- II. Web Server - תוכנת שרת המתקשרות בפרוטוקול HTTP האחראי על אחסון האתר.
- III. Database Server - מהוות את שכבת הנתונים. השרת אחראי על אחסון הנתונים של המערכת ומאפשר קריאה, הוספה, עריכה ומחיקה של נתונים בהתאם להרשאות ומקשור ל- Web Server אשר אחראי על איחסון ושליפת הנתונים.
- IV. Application Server - שכבת היישום שמבצעת את עיבוד הנתונים ושליטה בפונקציונליות. השרת אחראי על הפעלת האלגוריתמים והוא מציג את כל הנתונים לשכבה התצוגה של הלקוח.
- V. Google API - האלגוריתם מתקשך עם GOOGLE API בשני דרכים- גם על מנת להעלות ולמשוך קבצים מ Google Cloud Platform וגם על מנת לשלוח מייל דרך GMAIL. בכך לבצע שתי פעולות אלו, האלגוריתם משתמש ב- Credentials שהוגדרו מראש.

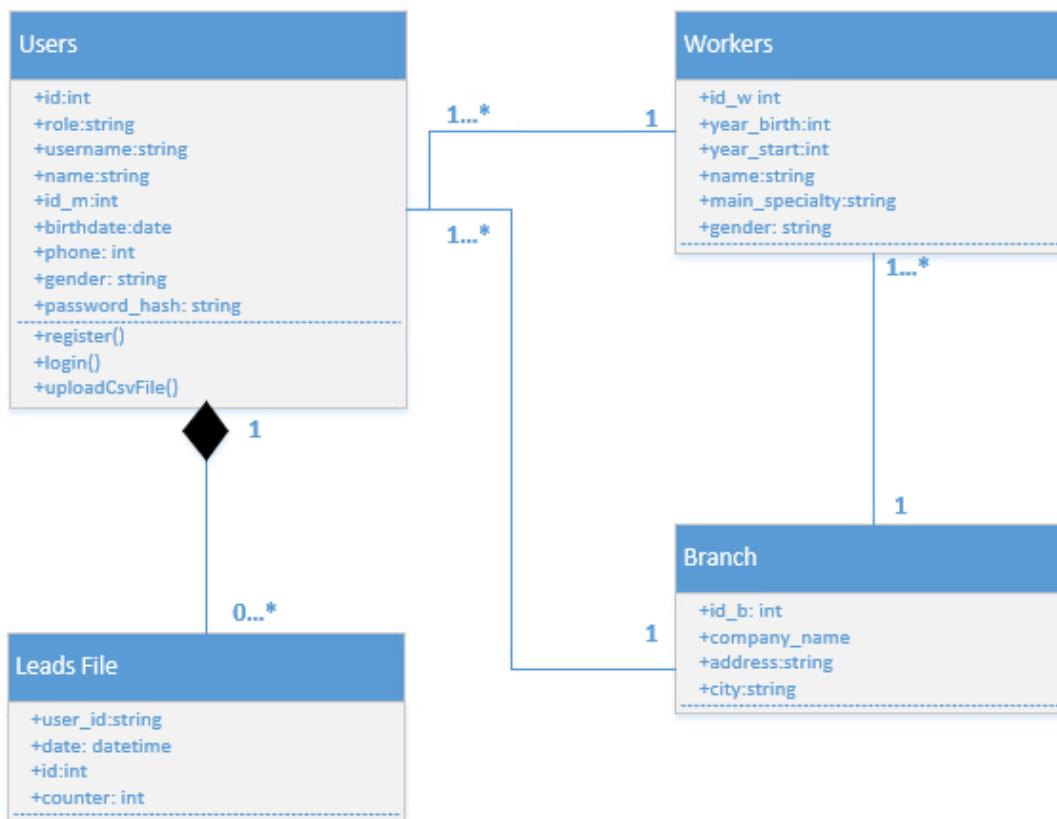
Component Diagram 8.2

בבסיס הנתונים של הדטה בייס של Lead Management יש 5 קומפוננטות עיקריות, שהם איש המכירות, מנהל הסניף, סניף וחברה, מתוך מנהל המכירות יוצא קומפוננטה של קובץ ניהול הלקוחים שגם הוא רשום כרשותה בדטה בייס שייה מעקב בכל פעם שהסניף משתמש במערכת שלנו, וזה מחובר אליו כי הוא הגורם בארגון שמעלה את הקובץ לאתר. מאותה סיבה מחובר קומפוננטה למנהל הסניף, כי לו יש גישה לדוחות ול-*Dashboard* של האתר.



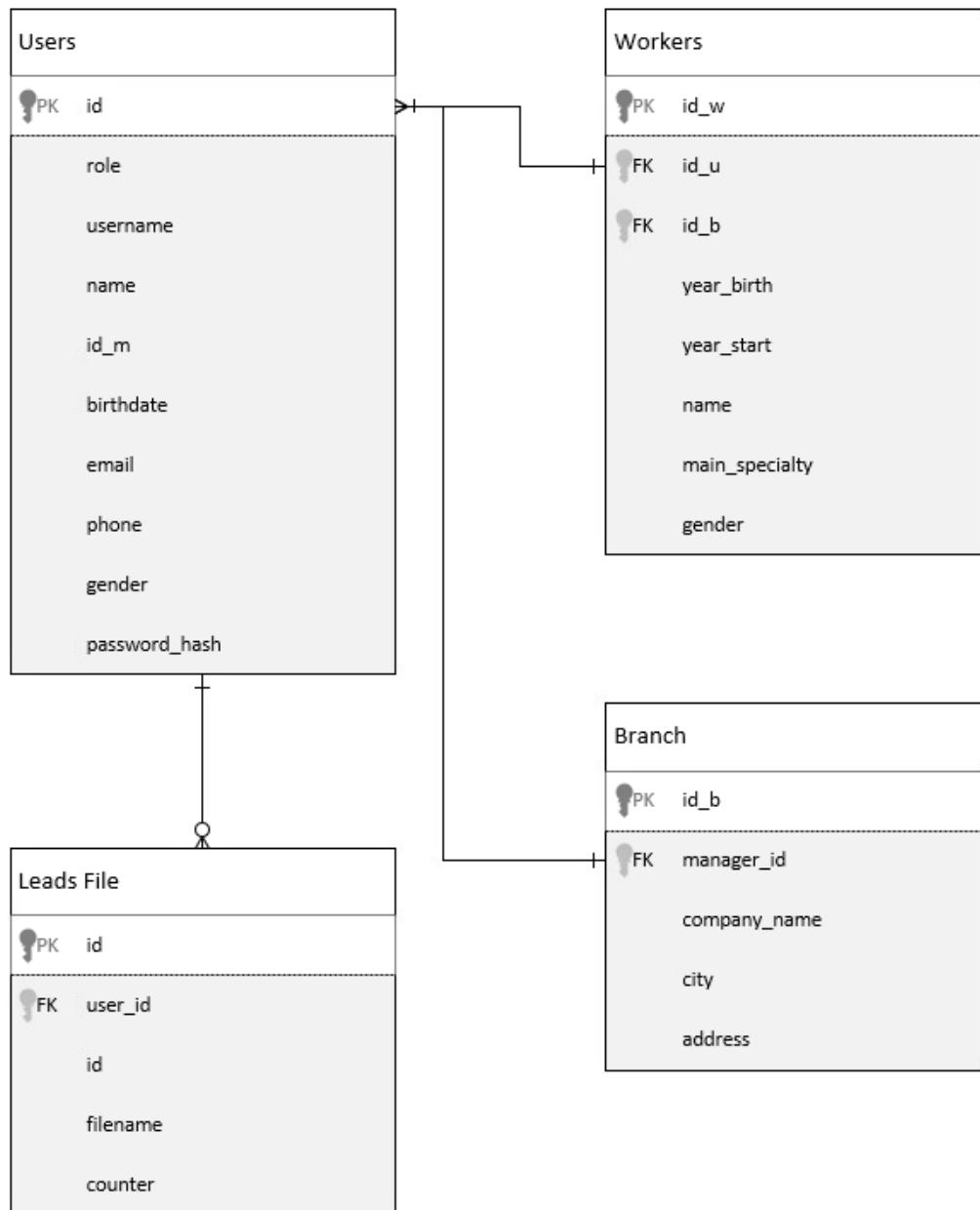
תרשים 8.2 Component Diagram .8.2

8.3 דיאגרמת מחלקות – Class Diagram



תרשים 8.3. דיאגרמת מחלקות

ERD – Entity Relationship Diagram 8.4



תרשים 8.4. תרשيم ERD

8.5 מיליון נתונים

8.5.1 טבלת Users

מפתח ראשי - סימנו בקוו תחתון את שם השדה
מפתח זר - בכוכבית לצד שם השדה.

مِلُون مُونَحَيْمٌ- طَبْلَة Users							
עריכים モトタリム	הסבר	אורך トビ	שדה コボハ	סוג ソガ	שם שדה	INT	Id
1-9999999	מספר סידורי אקראי של בעל התפקיד הנitinן באופן אוטומטי בעת הרשמה על ידי DB	6 תווים	לא - השדה לא חשוף למשתמש	STRING			
Branch Sales	בעת הרשמה על המשתמש בחומר תפקיד מתוך ChownDown-drilldown- ישנו רק שתי אפשרויות: או ניהול מכירות או ניהול סניף	עד 64 תווים	כן				role
אותיות, מספרים וSIMNIM	שם משתמש למערכת- חיב להיות ייחודי	עד 64 תווים	כן	STRING			username
אותיות וSIMNIM	שם פרטי ומשפחה	עד 64 תווים	כן	STRING			name
מספרים	תעודת זהות	-	כן	INT			id_m
בהתאם لتבנית התאריך	תאריך לידה - DD-MM-YY YYYY	תאריך	לא	DATE			birthdate
אותיות, מספרים וSIMNIM	כתובת מייל ייחודית	עד 64 תווים	כן	STRING			rmail
ספורות + וSIMNIM אם מעוניינים) בלבד	מספר טלפון	עד 64 תווים	כן	STRING			phone
Female, male	בעת הרשמה על המשתמש בחומר תפקיד מתוך ChownDown-drilldown- ישנו רק שתי אפשרויות: או זכר או נקבה	עד 64 תווים	כן	STRING			gender
אותיות, מספרים וSIMNIM	מוצפן- בעת הרשמה המרכיבת צפין ותישמר בDB סימנה מוצפנת	עד 128 תווים	כן	STRING			password_hash

טבלה 8.5.1. טבלת **Users**

 טבלת **Workers** 8.5.2

מילון מונחים - טבלת Workers						
ערכים מודולרים	הסבר	אורך	שדה חובה	סוג	שם שדה	
1-9999999	מספר סידורי אקראי של בעל התפקיד הנitin באופן אוטומטי בעט הרשמה על ידי DBn	6 תווים	לא- השדה לא חשוף למשתמש	INT	<u>id_w</u>	
1-9999999	מספר סידורי אקראי של בעל התפקיד הנitin באופן אוטומטי בעט הרשמה על ידי DBn	6 תווים	לא- השדה לא חשוף למשתמש	INT	branch_id*	
1-9999999	מספר סידורי אקראי של בעל התפקיד הנitin באופן אוטומטי בעט הרשמה על ידי DBn	6 תווים	לא- השדה לא חשוף למשתמש	INT	Id_u*	
אותיות, מספרים וסימנים	שנה שבה העובד התחל ל לעבוד בשבייל לחשב וותק	עד 64 תווים	C	STRING	year_birth	
אותיות, מספרים וסימנים	שנה שבה העובד התחל ל לעבוד בשבייל לחשב וותק	עד 64 תווים	C	STRING	year_start	
אותיות וסימנים	שם פרטי ומשפחה	עד 64 תווים	C	STRING	name	
אותיות וסימנים	התמחות ראשית	עד 64 תווים	C	STRING	main_specialty	
Female, male	בעט הרשמה על המשתמש לבחור תפקיד מתוך drilldown רק שניים: זכר או נקבה שתי אפשרויות: זכר או נקבה	עד 64 תווים	C	STRING	gender	

 טבלה 8.5.2. טבלת **Workers**

8.5.3 LeadsFile טבלת

מילון מונחים- טבלת LeadsFile							
שם שדה	סוג	שדה חובה	אורך	הסבר	ערכים מותרים		
id	INT	לא- השדה לא חשוף למשתמש	6 תווים	מספר סידורי אקראי של הקובץ הנוכחי באופן אוטומטי בעת העלאה של הקובץ ל- DB	1-9999999		
date	DATETIME	כן	14 תווים	תאריך ושעת ייצרת הקובץ hh:mm:ss DD/MM/YYYY	בהתאם לתבנית התאריך והשעה		
Filename	VARCHAR	לא	50 תווים	ניתן באופן אוטומטי בעת העילאת הקובץ על פי שם הקובץ	אותיות, סימנים ומספרים		
Counter	INT	לא	-	ניתן באופן אוטומטי בעת העילאת הקובץ על פי ספירה של מספר הקבצים של אלהו משתמש	מספרים		
user_id*	VARCHAR	לא- השדה לא חשוף למשתמש	6 תווים	שם המשתמש הנוכחי באופן אוטומטי בעת העלאה של הקובץ ל- DB	אותיות, סימנים ומספרים		

טבלה 8.5.3. טבלת LeadsFile

Branch 8.5.4

מילון מונחים - טבלת Branch						
ערבים מותרים	הסבר	אורן	שדה חוובה	סוג	שם שדה	
1-999999	מספר סידורי אקראי של הסניף הנוכחי באופן אוטומטי בעת הרשמה על ידי DB	6 תווים	לא- השדה לא חשוף למשתמש	INT	<u>b_id</u>	
1-999999	מספר סידורי אקראי של המשתמש הנוכחי באופן אוטומטי בעת הרשמה על ידי DB	6 תווים	לא- השדה לא חשוף למשתמש	INT	<u>manager_id</u> *	
אותיות תווים ומספרים	שם החברה	עד 64 תווים	cn	VARCHAR	<u>company_name</u>	
אותיות ותווים מקום	עיר שבה הסניף	עד 64 תווים	cn	VARCHAR	<u>city</u>	
אותיות, מספרים ותווים	כתובת של הסניף	עד 64 תווים	cn	VARCHAR	<u>address</u>	

טבלה 8.4. טבלת Branch

8.5.5 מילון נתונים – פירוט קבצי CSV בהם אנו עושים שימוש
הערה: סוגי המשתנים הוגדרו לפי שפת פיתון
8.5.5.1 קובץ הלידים המתקבל על ידי הליקון

מילון נתונים - קובץ הלידים המתתקבל על ידי הליקון						
ערכים מותרים	הסביר	אורך	שדה חובה	סוג	שם שדה	
1000000-9999999	כasher ליד נכון נכנס מתקובל על ידי הליקון הוא מקבל מספר סידורי	6 תווים	כן	INT	id_lead	
111111111-999999999	תעודת זהות של הליד	9 תווים	כן	INT	id	
True, False	האם הליד עסק או לא	5 תווים	כן	String	is_buisness	
אותיות	שם פרטי של הליקון	12 תווים	כן	String	first_name	
אותיות	שם אמצעי (אם יש) ושם משפחה של הליד	50 תווים	כן	String	last_name	
אותיות, מספרים וסימנים	כתובת מייל של הליד	50 תווים	כן	String	email	
Male, Female	מין	6	כן	String	gender	
0000-9999	שנת הלידה של הליקון	4	כן	INT	year_of_birth	
אותיות	ארץ המגורים של הליקון	50 תווים	כן	String	country	
אותיות ומספרים	כתובת המגורים של הליד	50 תווים	כן	String	address	
אותיות ומספרים	מקום העבודה של הליד	50 תווים	כן	String	company_name	
Phone, website, Instagram, Facebook, Google	הפלטפורמה בה נכון הליד למערכת	10 תווים	כן	string	platform	
אותיות ומספרים	תחום שבו עוסק הליד	20 תווים	לא	String	department	
אותיות	יצן הרכב המועדף על ידי הליקון	30 תווים	כן	String	car_type	
אותיות	מודל הרכב המועדף על ידי הליקון	30 תווים	לא	String	car_model	
משתנה מסוג date	תאריך יצירת הליד	10 תווים	כן	Date	creation_date	

1-24	שעת יצירת הליד	2 תווים	כ	INT	creation_time
1000-9999	שנת הרכב הרצויה	4 תווים	כ	INT	car_year
משתנה מסווג date	עד איזה תאריך הליד מעוניין בתקופת ההשכרה	10 תווים	כ	date	rental_period

טבלה 8.5.5.1 קובץ הלידים המתקבל על ידי הלוקו

8.5.5.2 קובץ הנתונים לאחר שלב ה-ETL

מילון נתונים - סט הנתונים לאחר שלב ה-ETL						
ערכים מותרים	הסבר	אורך	שדה	חובב	סוג	שם שדה
100000-999999	כasher ליד נכנס מתקובל על ידי הליקות הוא מקבל מספר סידורי	6 תווים	C	INT		id_lead
11111111-99999999	תעודת הזהות של הליד	9 תווים	C	INT		id
True, False	האם הליד עסק או לא	5 תווים	C	String		is_buisness
אותיות	שם פרטי של הליקות	12 תווים	C	String		first_name
אותיות	שם אמצעי (אם יש) ושם משפחה של הליד	50 תווים	C	String		last_name
אותיות, מספרים וסימנים	כתובת המייל של הליד	50 תווים	C	String		email
Male, Female	מין	6	C	String		gender
0000-9999	שנת הלידה של הליקות	4	C	INT		year_of_birth
אותיות	ארץ המגורים של הליקות	50 תווים	C	String		country
אותיות ומספרים	כתובת המגורים של הליד	50 תווים	C	String		address
אותיות ומספרים	מקום העבודה של הליד	50 תווים	C	String		company_name
Phone, website, Instagram, Facebook, Google	הפלטפורמה בה נכנס הליד למערכת	10 תווים	C	string		platform
אותיות ומספרים	תחום שבו עסוק הליד	20 תווים	לא	String		department
אותיות	ישן הרכב המועדף על ידי הליקות	30 תווים	C	String		car_type
אותיות	מודל הרכב המועדף על ידי הליקות	30 תווים	לא	String		car_model
משתנה מסווג date	תאריך יצירת הליד	10 תווים	C	Date		creation_date

1-24	שעת יצירת הליד	2 תווים	ס	INT	creation_time
1000-9999	שנת הרכב הרצואה	4 תווים	ס	INT	car_year
משתנה מסווג date	עד איזה תאריך הליד מעוניין בתקופת ההשכרה	10 תווים	ס	date	rental_period
	מה הגיל המספרי של הליד		מחוש ב	INT	age
	מחיר המחרiron של הרכב המבוקש	64 ביטים	ס	Float	car_price
	תקופת ההשכרה הרצואה בימים	1000 תווים	ס	INT	desirable_renta l_days
	משתנה קטגוריאלי באיזה פרק זמן ביום פנה אלינו הליד	עד 20 תווים	ס	String	time_catagor
	רווח החברה שבה מועסק הליד	64 ביטים	מחוש ב	Float	profit
	שווי השוק שבו מועסק הליד	64 ביטים	מחוש ב	Float	Market Cap

טבלה 8.5.5.2. קובץ הנתונים לאחר שלב ה-ETL

8.5.5.3 מאגר המידע של הרכבים

מילון נתונים - מאגר המידע של הרכבים						
ערכים מותרים	הסבר	אורך	שדה חויה	סוג	שם שדה	
1000000000-9999999999	מספר זהה ייחודי של כל רכב	10 תווים	כן	INT	id_car	
מוגבל בזיכרון	מחיר הרכב	64 ביטים	כן	Float	price	
1000-9999	שנת הייצור של הרכב	4	כן	String	year	
אותיות ומספרים	Nazan הרכב	50 תווים	כן	String	car	
אותיות ומספרים	מודל הרכב	50 תווים	כן	String	model	
אותיות, מספרים	סוג הצילינדר של הרכב	50 תווים	לא	String	cylinders	
אותיות	סוג הדלק המotor ברכב	50 תווים	לא	String	fuel	

טבלה 8.5.5.3. מאגר המידע של הרכבים

8.5.5.4 מאגר המידע של החברות העסקיות

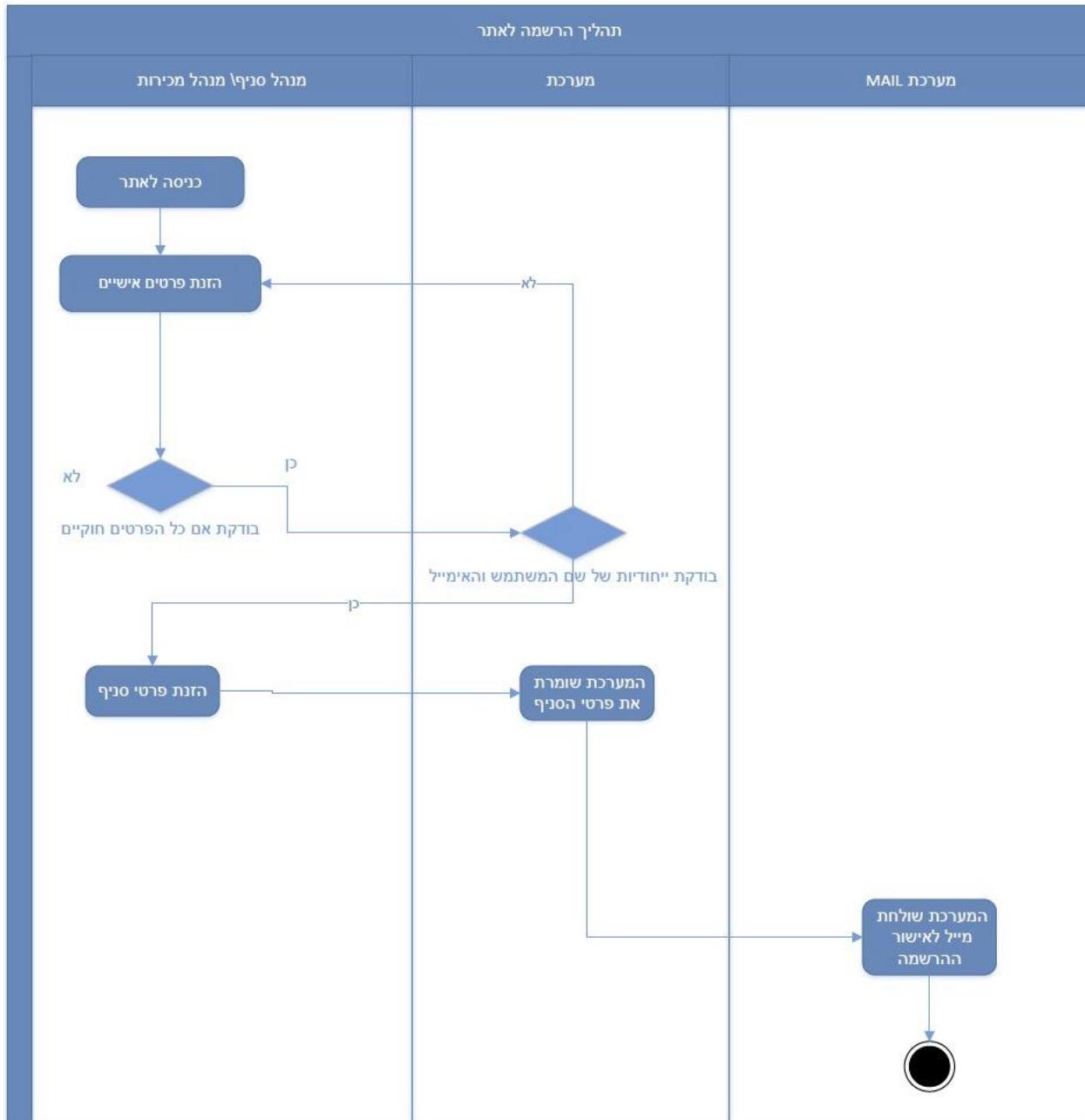
מילון נתונים - מאגר המידע של החברות העסקיות						
ערכים מותרים	הסבר	אורך	שדה חובב	סוג	שם שדה	
10000-99999	מספר מזהה ייחודי של כל חברת	50 תווים	ק	INT	id_company	
111111111-999999999	תעודת זהות של הליד	50 תווים	ק	String	company	
מוגבל בזיכרון	דירוג החברה מבחינת שווי שוק	64 ביטים	ק	int	rank	
מוגבל בזיכרון	שינוי הדירוג של חברות מבחינת שווי שוק לעומת שנה שעברה	64 ביטים	ק	int	rank_change	
מוגבל בזיכרון	הכנסות החברה	64 ביטים	ק	Float	revenue	
מוגבל בזיכרון	רווח החברה	64 ביטים	ק	Float	profit	
מוגבל בזיכרון	כמות העובדים בחברה	64 ביטים	ק	Int	num. of employees	
אותיות	מגזר העסק שבו עסקת החברה	50 תווים	String	String	sector	
אותיות	העיר שבו ממוקמת הסניף הראשי של החברה	עד 50 תווים	ק	String	city	
אותיות	מדינה שבה ממוקמת הסניף הראשי של החברה	עד 50 תווים	ק	String	state	
Yes, no	האם החברה חדשה	3 תווים	לא	String	new_comer	
Yes, no	האם מקים החברה הוא ذכר	3 תווים	לא	String	ceo_founder	
Yes, no	האם מקימת החברה היא נקבה	3 תווים	לא	String	ceo_women	
Yes, no	האם החברה מרווחת	3 תווים	לא	String	profitable	
מוגבל בזיכרון	הדרוג השנתי הקודם של החברה	64 ביטים	לא	Int	prev_rank	

50 תווים	שם מנכ"ל החברה	50 תווים	לא	String	ceo
אותיות ומספרים	קישור לאתר החברה	100 תווים	לא	String	website
מוגבל בזיכרון	שווי שוק החברה	64 BITSIM	כן	Float	Market Cap

טבלה 8.5.5.4 מאגר המידע של החברות העסקיות

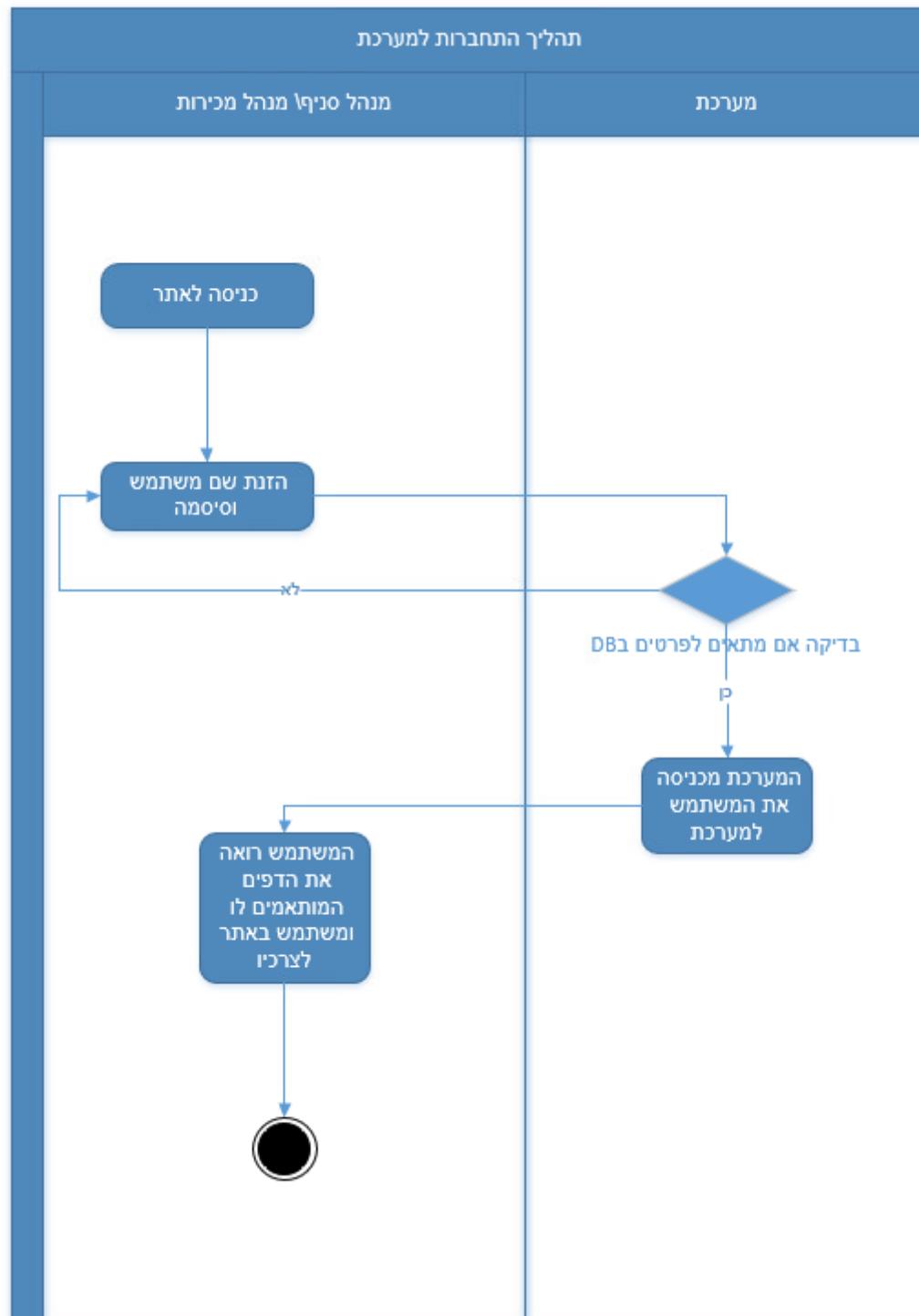
8.6 דיאגרמת רצף או תרשيم פעילות לתהליכיים העיקריים במערכת

8.6.1 תהליך הרשמה לאתר



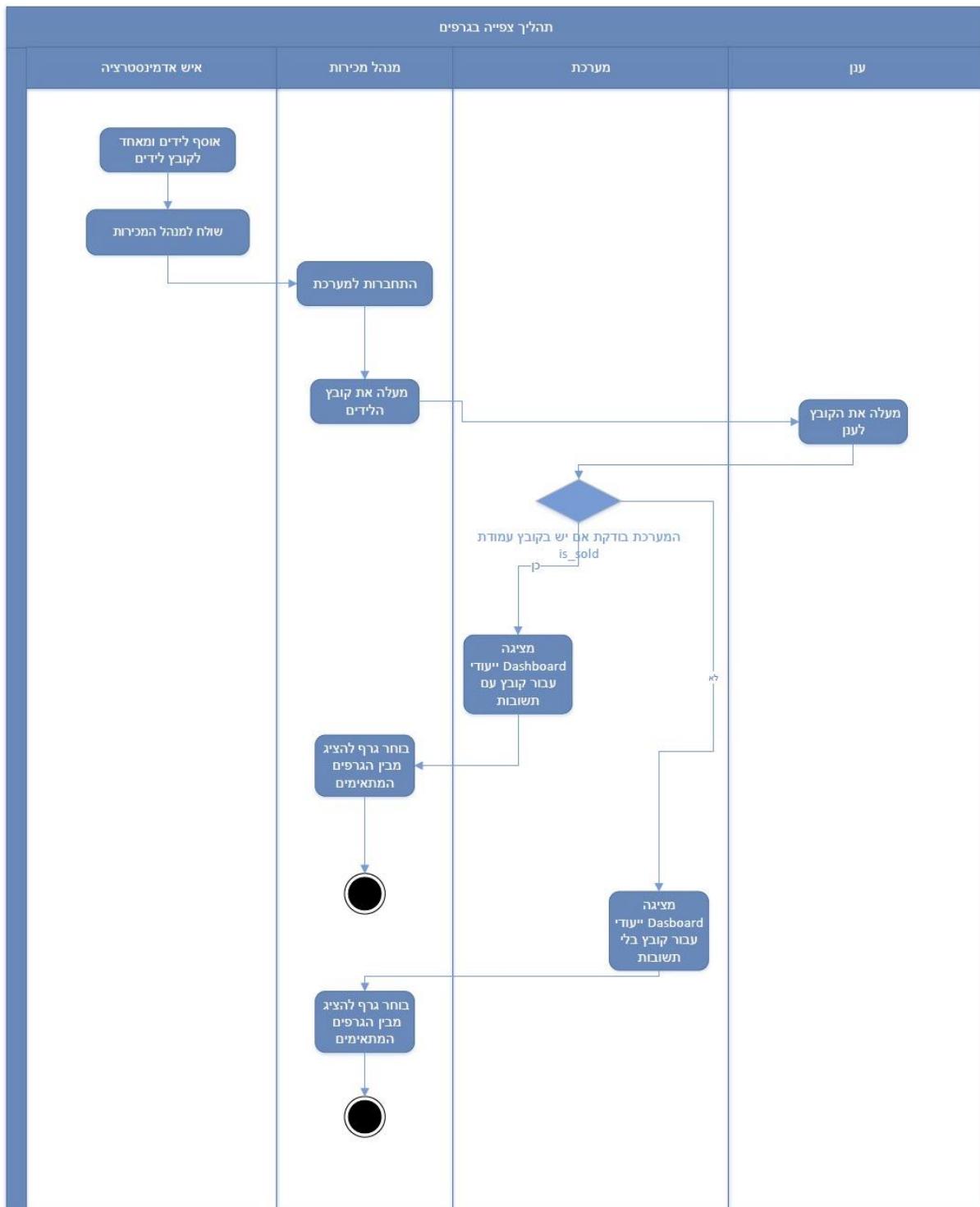
תרשים 8.6.1 – תהליך הרשמה לאתר – Activity Diagram

8.6.2 תהליך התחברות למערכת



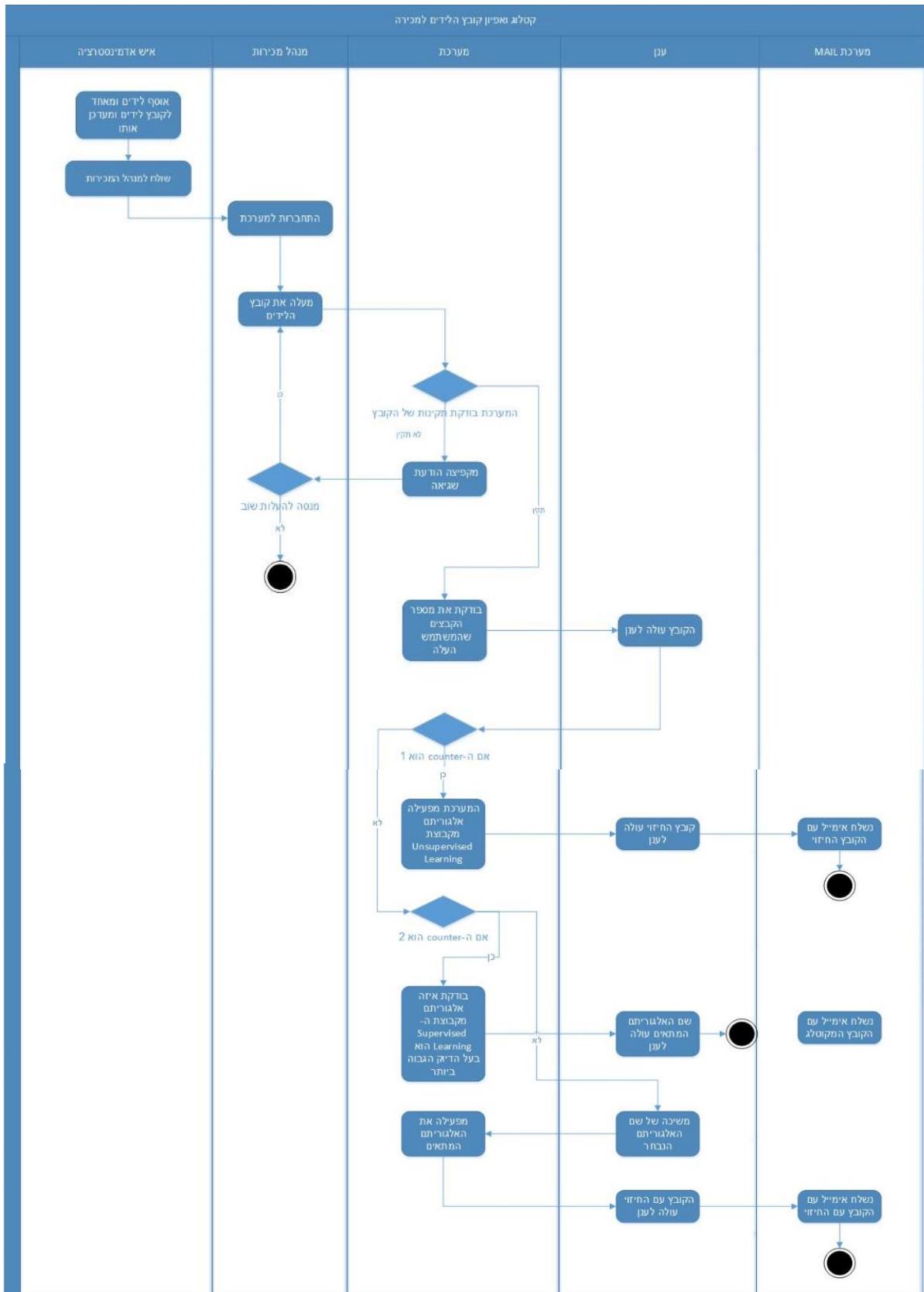
תרשים 8.6.2 – תהליך התחברות

8.6.3 תהליך צפיה בגרפים

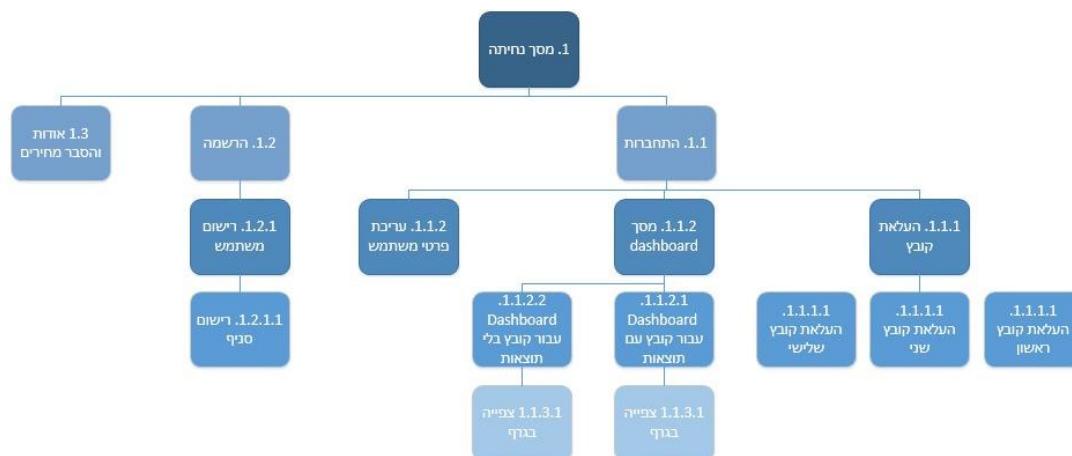


תרשים 8.6.3 – תהליך צפיה בגרפים

8.6.4 תהליכי העלאת קובץ לידיים



תרשים 8.6.4 – תחילת העלאת קובץ לידיים

תרשים 8.7 תרשימים עץ המרכיבים

תרשים 8.7. עץ המרכיבים

9 תיאור של אלגוריתמים ותהליכי חישוב שמבצעת המערכת

K - Means Clustering 9.1

אלגוריתם K-means Clustering – K עובד באופן הבא:

- I. בחרת המספר K המיצג את מספר האשכולות.
- II. נבחרו באקראיות K נקודות. ככלומר נבחרות באופן אקראי K נקודות, שהן יהיו האשכולות הראשונים (לאחר מכן האשכול המרכזי ישתנה).
- III. מציאת המרחק האוקלידי של כל נקודה במרחב הנתונים של האשכול (K).
- IV. הקצתה כל נקודה נתונים למרצט הקרוב ביותר באמצעות המרחק שנמצא בשלב הקודם.
- V. חוזה על 2 עד 4 עבור מספר קבוע קבוע של האיטרציות או עד שהמרכזים לא ישתנו.
- VI. סיום המודול

9.1.1 נוסחה מתמטית

נניח שיש לנו קבוצה של תכיפות, (x_1, \dots, x_n) כאשר כל תכיפות היא וקטור ממשי יכול להיות בעל מספר מדדים (כמו במרקלה שלנו שלכל נקוח X יש כמה Attributes לכך הוא רב ממד'). מטרת המודל היא לחלק את ח התכיפות לא אשכולות, על מנת למזער את סכום המרחקים בין התכיפות בתוך האשכול.

אפשר גם להסתכל על כך שהאלגוריתם רוצה להכניס ח תכיפות (האיקסים) לתוך סט של $S = \{S_1, \dots, S_K\}$ כדי למזער את סכום הריבועים בתוך אשכול.

9.1.2 מרחק אוקלידי בין שתי נקודות במרחב:

באמצעות המרחק האוקלידי מחשבים את המרחק בין כל נקודה נתונם לאשכול מרכזים באמצעות הנוסחה **כדי**

$$d(\mathbf{p}, \mathbf{q}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2}.$$

נוסחה 5.1.3.1.1. מרחק אוקלידי

נקודות הנתונים מוקנית למרכז האשכול שהマーク שלו ממרכז האשכול הוא מינימום מכל מרכז'

9.1.3 הקצתה כל נקודה לאשכול הקרוב ביותר:

$$c_i = \frac{1}{|S_i|} \sum_{x_i \in S_i} x_i$$

נוסחה 5.1.3.1.3. הקצתה כל נקודה לאשכול הקרוב ביותר

כאשר x_i הוא קבוצת כל הנקודות שהוקטו לאשכול ה- i , ככלומר גודל הקבוצה בתוך ה-Cluster.

K - Prototypes Algorithm 9.2

המודל אשר פותח על ידי הוואנג (קישור למאמר בסעיף 16) יצר שילוב בין שני המודלים: - K Means Clustering ו- K-Modes Clustering. ועל ידי זה, אפשר לבצע ניתוח אשכולות עבור מידע הכלל מעמודות קטגוריאליות ונומריות.

האלגוריתם היברידי מכיל שילוב של המוצעים מהעמודות הנומריות והערך השכיח מהעמודות הקטגוריאליות.

9.2.1 שלבי המודל:

- I. K איברים מהדата נבחרים באופן רנדומלי והם מייצגים את מרכז האשכולות ההתחלתיים. K הינו מספר שהוגדר מראש.
- II. המודל משתמש במקדם שונות המורכב מחיבור של מרחק המיניג (עבור העמודות הקטגוריאליות) ושורש של המרחק האוקלידי (עבור העמודות הנומריות) עבור X הпромופרציה של שני המרחקים מותאמת על ידי הפרמטר גמא ותפקידו לשלוט על המשקל היחסי בין שני המודלים. בסוף החישוב X משתמש למרכז האשכול הקרוב ביותר אליו.

$$d(x_i, q_l) = \gamma \sum_{s=1}^p \delta(x_{i,s}^C - q_{l,s}^C) + \sum_{s=p+1}^m \sqrt{(x_{i,s}^N - q_{l,s}^N)^2},$$

$$\text{where } \delta(x_{i,s}, q_{l,s}) = \begin{cases} 0, & x_{i,s} = q_{l,s}, \\ 1, & x_{i,s} \neq q_{l,s}. \end{cases}$$

נוסחה 9.2.1. פונקציית ההפסד עבור K-Prototypes

- d - המשתנים הנומריים
- p - המשתנים הקטוריים
- X_i - הנקודה החדשה המוחשבת שעל המודל לשיר למרכז אשכול. בהתאם למרכז האשכול החדש, מחשבים את מקדם השונות שוב. עדכן מחדש של מרכז האשכולות.
- III.
- IV.
- V.
- VI.
- VII.
- חישוב פונקציית העלות אותה המודל רוצה לצמצם ככל שאפשר:

$$F(U, Q) = \sum_{l=1}^k \sum_{i=1}^n u_{il} d(x_i, q_l).$$

- VI.
- VII. המודל יחזיר על סעיפים 2,3,4,5 עד שפונקציית העלות לא תשתנה יותר, במידה והיא לא תשתנה ניתן להגיד שהגענו למודל עם מרכזי אשכולות האידאליים עבור הדטה עבור K (מספר האשכולות) שנבחר.

9.3 רגסיה לוגיסטי – Logistic Regression

בchner בסקירת הספרות בסעיף [2.2.4](#) את האלגוריתם באופן כללי, אך בסעיף זה נתמקד בהסביר האלגוריתם מבחינה מתמטית.

רגסיה לוגיסטי יכולה להיאזר / לנבוע מאנלוגיה להשערה הרגסיה היליניארית שהיא:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \quad \text{where} \quad \epsilon_i \sim^{iid} N(0, \sigma^2)$$

נוסחה 2.5.1.3.2 - רגסיה ליניארית

X – האיקסים

β – וקטור הפרמטרים של המקדמים

\in – השגיאה

וניתן לכתוב זאת גם כך בצורה מטריצונית:

$$h_{\theta}(\mathbf{x}) = \theta^T \mathbf{x}$$

כאשר וקטור טהה הוא וקטור המקדמים, טטה בחזקת T מייצגת transpose, והאיקסים מייצגים את וקטור הערכים.

השערה הרגסיה הלוגיסטי נבנית מהשערה הרגסיה היליניארית בכך שהיא משתמש בפונקציה הלוגיסטי:

$$h_{\theta}(\mathbf{x}) = g(\theta^T \mathbf{x})$$

$$g(z) = \frac{1}{1 + e^{-z}}$$

התוצאה היא השערת הרגסיה הלוגיסטי:

$$H_{\theta}(X) = \frac{1}{1 + e^{-\theta^T x}}$$

הפונקציה (z) g היא הפונקציה הלוגיסטי, הידועה גם כפונקציה הסיגמודידית. לפונקציה הלוגיסטי יש אסימפטוטיים ב-0 ו-1, והוא חוצה את ציר ה- y ב-0.5.

היתרון ברגסיה הלוגיסטי שהוא יכול לבצע סיווג ביןארי בצורה מהירה ויעילה.

הסיבה לכך היא ש"ספ" הרגסיה הלוגיסטי מוגדרת ל- $0.5 = (z)g$, כאשר נכנס את הפרמטרים של הליד, והפלט של הפונקציה ייתן לנו ערכים שהם מעל 0.5, נוכל לסווג אותם כ-1 כלומר הליד ימוך, ולהפוך, אם נקבל ערכים מתחת לחצי, נוכל לסווג אותם כ-0, כלומר הליד סוג בעל סיכוי נמוך להפוך למוכר.

9.4 עץ החלטה – Decision Tree

סקרנו בסקירת הספרות בסעיף [3.5.2](#) את האלגוריתם באופן כללי, لكن בסעיף זה נתמקד בבחינת האלגוריתם מבחינה מתמטית.

באופן כללי-ב-Decision Trees, ב כדי לנבأ תווייה מחלוקת (Labels) לרשומה – אנחנו נתחיל את התהיליך משורש העץ (Root), אנו משווים את העריכים של תוכנת השורש עם התוכונה של הרשומות, ועל בסיס השוואה, אנו עוקבים אחר הענף המתאים לערך זה ועוברים לצומת הבא.

ישנים 2 סוגים עיקריים של עציים:

א. עץ החלטות משתנה קטגוריו

ב. עץ ההחלטה אשר מנבה משתנה יעד רציף (כלומר נומי)

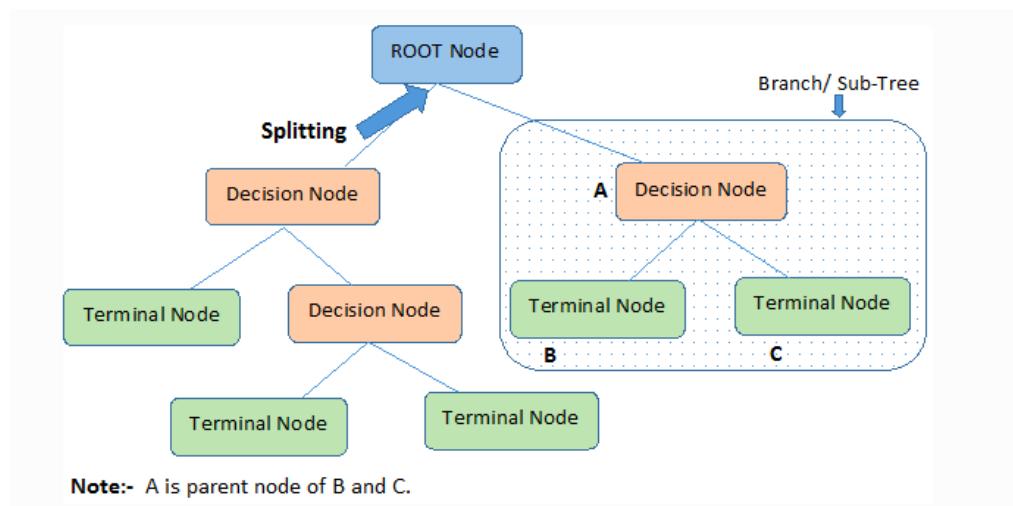
כמובן שבפרויקט נתמקד בעץ ההחלטה משתנה קטגוריו.

9.4.1 טרמינולוגיה הקשורה לעצי החלטה

הסבר	מונה
מייצג את כל האוכלוסייה או המדגם אשר אחריו המדגם מחולק לשניים או יותר.	צומת שורש – Root Node
זהו תהיליך של חלוקת צומת לשני צמות משנה או יותר.	חילוק - Splitting
כאשר צומת משנה מתפצל לצמות משנה נוספים, הוא נקרא צומת ההחלטה.	צומת ההחלטה – Decision Node
צמותים אינם מפוצלים נקראים עלה או צומת מסוף.	עליה / צומת מסוף - Leaf / Terminal Node
כאשר אנו מסירים צמות משנה של צומת ההחלטה תהיליך זה נקרא גיזום. אפשר לומר לומר תהיליך הפוך של פיצול.	גזם – Pruning
ענף משנה של העץ יכול נקרא ענף או עץ משנה.	ענף / תת-עץ - Branch / Sub-Tree
צומת המחולק לצמות משנה נקרא צומת אב של צמות אב. צומת אב וצאצא - parent and child node	צומת אב וצאצא - parent and child node
הוא מدد עד כמה חילוקים (Splits) העץ החלטה יכול לקיים לפני שהוא מבצע סיווג.	עומק - Depth

טבלה 9.4.1. טבלת טרמינולוגיה עצי ההחלטה

דוגמה למונחי הטרמינולוגיה של עצי החלטה:



איור 4. מדגים את הטרמינולוגיה של עצי החלטה

ניתן להפעיל את עצי ההחלטה, או יותר נכון תהליכי הבחירה של ה-Attributes על פי כמה אלגוריתמי בחירה שונים, אך נבחר להתמקד ב-ID3.

9.4.2 ID3 Algorithm

האלגוריתם ID3 בניית עצי החלטה תוך שימוש בגישה חיפוש "חמדנית" מלמעלה למטה דרך מרחב הענפים האפשריים ללא חזרה לאחור. אלגוריתם חמדן, כפי שהוא מגדן, תמיד עושה את הבחירה שנראית הטובה ביותר באותו רגע.

9.4.3 השלבים באלגוריתם ID3:

- I. מתחילה עם S הסט המקורי כזומת הבסיס.
- II. בכל איטרציה של האלגוריתם, הוא מחשב ל-attribute של קבוצת S את האנטרופיה (H) או Information Gain של תוכנה זו.
- III. בחירת התוכנה שיש לה את האנטרופיה הגדולה ביותר או את רוח המידע הגדול ביותר.
- IV. לאחר מכן, הסט S מפוצלת על-ידי התוכנה שנבחרה כדי להפיק קבוצה חדשה של הנזונים.
- V. האלגוריתם ממשיך לחזור על עצמו בכל קבוצה משנה בהתחשב רק בתוכנות שלא נבחרו קודם, ככלומר שלא חולקו כבר.

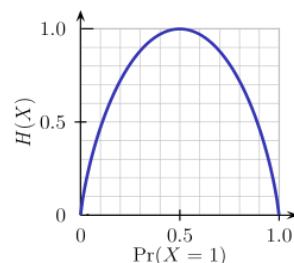
9.4.4 אמצעי בחירת תוכנות

אם מערכת הנתונים מורכב מ-N תוכנות אז ההחלטה איזו תוכנה למקם בשורש או ברמות שונות של העץ צמצמים פנימיים היא שלב מורכב. רק בחירה אקראיית של כל צומת להיות השורש לא יכולה לפתור את הבעיה. אם נפעל בגישה אקראית, היא עלולה לתת לנו תוצאות עם דיקון נמור.

לפתרון בעיית בחירת התוכנות הזה, חוקרים עבדו והגו כמה פתרונות ואנו נתמקד ב-2 מונחים בשבייל לפתרות את בעיה זו ב-אנטרופיה ו-Information Gain.

9.4.5 אנטרופיה – Entropy

אנטרופיה היא ממד לאקראיות במידע הקיימים שלנו (כלומר במידע סט). ככל שהאנטרופיה גבוהה יותר (כאשר אנטרופיה היא ממד בין 0 ל-1), קשה יותר להסיק מסקנות מהמידע זהה. הטלת מטבע היא דוגמה לפעולה המספקת מידע אקראי.



graf 9.4.5. אנטרופיה

מהגרף שלמעלה, די ברור שהאנטרופיה (H) היא אפס כאשר ההסתברות היא או 0 או 1 כלומר אנחנו בטוחים לגבי הסיגו שלנו.

האנטרופיה מקבלת ערך מקסימלי כאשר ההסתברות היא 0.5 כאשר היא מציגה אקראיות מושלמת בנתונים ואין סיכוי שנוכל לקבוע בצורה מושלמת את התוצאה.

איך שיטה זו מתחברת לאלגוריתם ID3?

ID3 עוקב אחר הכלל - ענף עם אנטרופיה של אפס הוא צומת עלה, וענף עם אנטרופיה יותר מ一封 ציר פיצול נוסף.

9.4.6 הנוסחה של אנטרופיה מחושבת בצורה הבאה:

$$E(S) = \sum_{i=1}^c -p_i \log_2 p_i$$

נוסחה 5.1.3.3.6. אנטרופיה

כאשר:

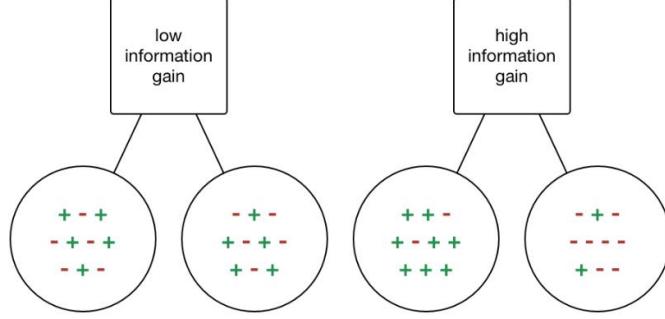
S – מצב הנוכחי.

i – הסתברות לאירוע i של מצב S
Information Gain – רוח מידע

9.4.7 רוח מידע – Gain

רוח מידע או GI הוא מאפיין סטטיסטי המודד עד כמה תכונה נתונה מפרידה בין דוגמאות האימון לפי סיווג המידע שלהם.

בנית עץ החלטות מתקדת במצבת תכונה שמחזירה את רוח המידע הגבוה ביותר ואת האנטרופיה הקטנה ביותר.



איור 9.4.7. רוח מידע

רוח מידע הוא ירידה באנטרופיה. הוא מחשב את ההבדל בין אנטרופיה לפני הפיזול לבין אנטרופיה המוצעת לאחר הפיזול של מערך הנתונים בהתאם על ערכי התכונות (Attributes) הנתונים.

אלגוריתם עץ החלטות ID3 משתמש ברוח מידע.

9.4.7.1 מבחינה מתמטית GI מחושב כך:

$$IG(T, X) = Entropy(T) - Entropy(T, X)$$

או במילים אחרות:

$$IG = Entropy(Before) - \sum_{j=1}^K Entropy(j, after)$$

כאשר "לפני" זה הדאטה סט לפני החילוק, K - הוא מספר קבוצות המשנה שנוצרו על ידי הפייזול, ו- $(j, after)$ היא תת-קבוצה j לאחר הפייזול.

9.4.8 ההבדל בין Random Forest לעצ החלטה

על ההבדלים בין Random Forest לעצ החלטה פירטנו בסעיף הבא [9.5.4](#).

Random Forest 9.5

"עיר אקראי" הוא אלגוריתם למידת מכונה ממשפחת הלמידה המפוקחת שנמצא בשימוש נרחב בעיות סיווג ורגסיה. האלגוריתם בונה עצי החלטה על מדגים שונים ומשתמש בשכיח למקרי סיווג, ובמוצע במקרה של רגסיה.

אחד המאפיינים החשובים ביותר של אלגוריתם העיר אקראי (כמו בעצ החלטה) שהוא יכול להתמודד עם נתונים המכילים משתנים רציפים כמו במקרה של רגסיה ומשתנים קטגוריאליים כמו במקרה של סיווג.

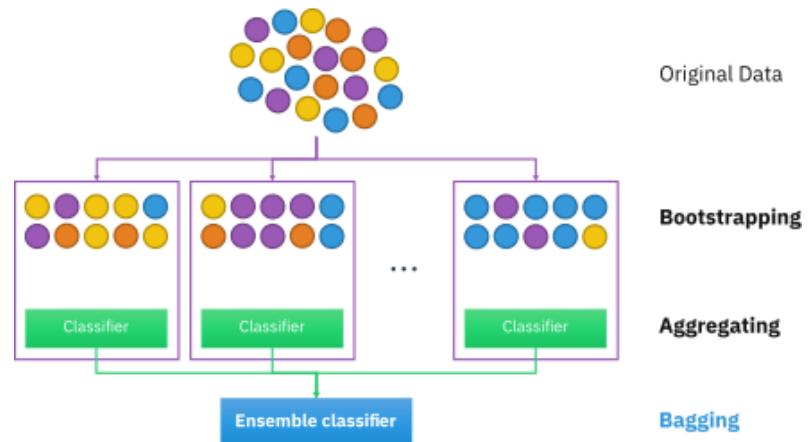
בשביל לבחון את האלגוריתם Random Forest נפרט ראשית על כמה מונחים בסיסיים.
Ensemble (מכול) - שילוב של מספר מודלים. לעומת זאת של מודלים המשמשים לביצוע תחזיות:
Boosting Bagging Ensemble – משתמש ב-2 שיטות [Boosting](#) ו-[Bagging](#)

Boosting Bagging 9.5.1

Bagging, הידוע גם בשם [Bootstrap Aggregation](#) היא טכניקת ה-[Ensemble](#) המשמשת את העיר אקראי.

Bagging בוחר מדגם אקראי ממערך הנתונים. لكن כל מדגם נוצר מהדגימות (דגימות Bootstrap) המוסףות על ידי הנתונים המקוריים עם החלפה המכונה [Row Sampling](#).
שלב זה של דגימת שורה עם החלפה נקרא [bootstrap](#).Cut כל מודל מאומן באופן עצמאי ובסופו של דבר מביא תוצאות (פלט).

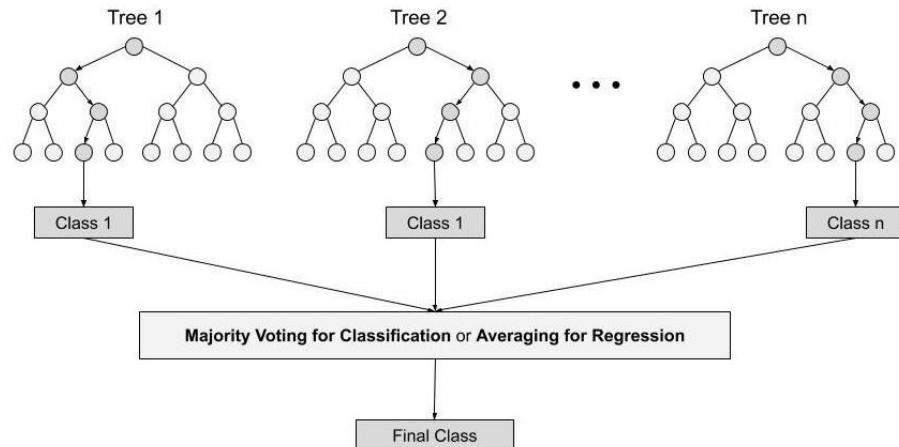
הפלט הסופי מבוסס על הצבעת הרוב לאחר שילוב התוצאות של כל המדגים. שלב זה הכלול שילוב של כל התוצאות ויצירת פلت המבוסס על הצבעת רוב מכונה [Aggregation](#) כלומר צבירה.



איור 9.5.1. מודים את תהליך Bagging שמבצע העיר אקראי

9.5.2 שלבי המודל

- I. נלקח מספר k של רשומות אקראיות ממערך הנתונים הכלול A מספר רשומות.
- II. עצי החלטה עצמאיים (כלומר כל עץ לעצמו) נבנים עבור כל מדגם.
- III. כל עץ החלטה יפיק פלט
- IV. הפלט הסופי יחושב על סמך הצבעת הרוב או ממוצע עבור סיווג ורגסיה בהתאם.



תרשים 9.5.2. דוגמה לביצוע והחלטה של יער אקראי

9.5.3 מאפיינים חשובים של Random Forest

- I. לא כל התכונות והערכים נחשבים ביצירת עץ בודד, כל עץ שונה מקודמו.
- II. כל עץ נוצר באופן עצמאי מ选出 הערכים וה-Features שהכנסנו כקלט, ולכן, איןנו יכולים למשות שימוש מלא במעבד כדי לבנות יתרות אקראיים (פערולה הצורכת CPU גבוהה).
- III. ביער אקראי לא נתבקש להפריד את הנתונים לדאטה סט אימון ודאטה סט מבחן מכיוון שתמיד יהיו 30% מהנתונים שלא'Rאים בעץ ההחלטה (מכיוון שכל עץ נוצר באופן עצמאי).
- IV. הפלט של המודל נחשב יציב מכיוון שההתוצאה מבוססת על שכיח או על ממוצע (בהתאם לנ נתונים).

9.5.4 ההבדל בין עץ החלטה ל- Random Forest

חשוב לציין את ההבדלים העיקריים של עץ החלטה רגיל-ו-Random Forest שהם פומים מושגים אלו ונשמעים אותו הדבר.

עץ החלטה	יער אקראי
עץ החלטה לרוב סובלילס-Over Fitting מכיוון שעש' יכול להיות גדול ללא "שליטה" (כלומר מספר הענפים אינו מוגבל).	יתרונות אקראיים נוצרים ממתת-קבוצות של נתונים והפלט הסופי מבוססת על ממוצע או שכיח מה שגורם ליתרונו על בעיית ה-Over Fitting.
עץ החלטה יחיד הרבה יותר מהיר לחישוב	יער אקראי יותר איטי
כאשר מערכ נתונים עם תכונות נלקח כקלט על ידי עץ החלטה, הוא יגבש סט כלליים לביצוע חיזוי.	יער אקראי בוחר באקראי תוצאות, בונה עץ החלטות ונלקחת התוצאה הממוצעת. لكن הוא לא משתמש אף קבוצה של נוסחים.

טבלה 9.5.4. הבדלים בין עץ החלטה ל- Random Forest

Grid Search CV 9.6

במהלך הפרויקט אנחנו משתמשים בשיטה שנקראת Grid Search CV על עץ החלטה על מנת למצוא את ההיפר פרמטרים הטובים ביותר שבאמצעותם נבנה את העץ בעל אחוז הדיקט הגבוה ביותר.

בשביל להבין את מטרת השיטה קודם נבון מה זה היפר פרמטרים.

9.6.1 היפר פרמטרים

למודל למידת מכונה יש מספר פרמטרים שאינם מוגדרים מראש והם צריכים להיות מוגדרים על ידי מדען הנוטנים או המהנדס.

פרמטרים אלו שלוטים על דיקוק המודל. לכן, היפר פרמטרים חשובים במיוחד בפרויקט מדעי נתונים וקרית מידע.

לדוגמה, קצב הלמידה של רשת נוירונים הוא היפר פרמטר מסוון שהוא קבוע על ידי המהנדס לפני שננתנו הקלט מתוקבלים במודל, או במקורה שלנו, שינוינו את היפר פרמטרים של עץ החלטה כמו עומק המקסימלי של העץ, כמות העלים המינימלית לפני שנבצע חילוק, ומהו מינימום מקבץ השורות שצורך להתקבל בשוביל להפוך לעלה.

הרחבת הנושא של הטרמינולוגיה של עץ החלטה מורחבת בסעיף [9.4.1](#).

از מה למשה עושה טכניקת "חיפוש רשת"?

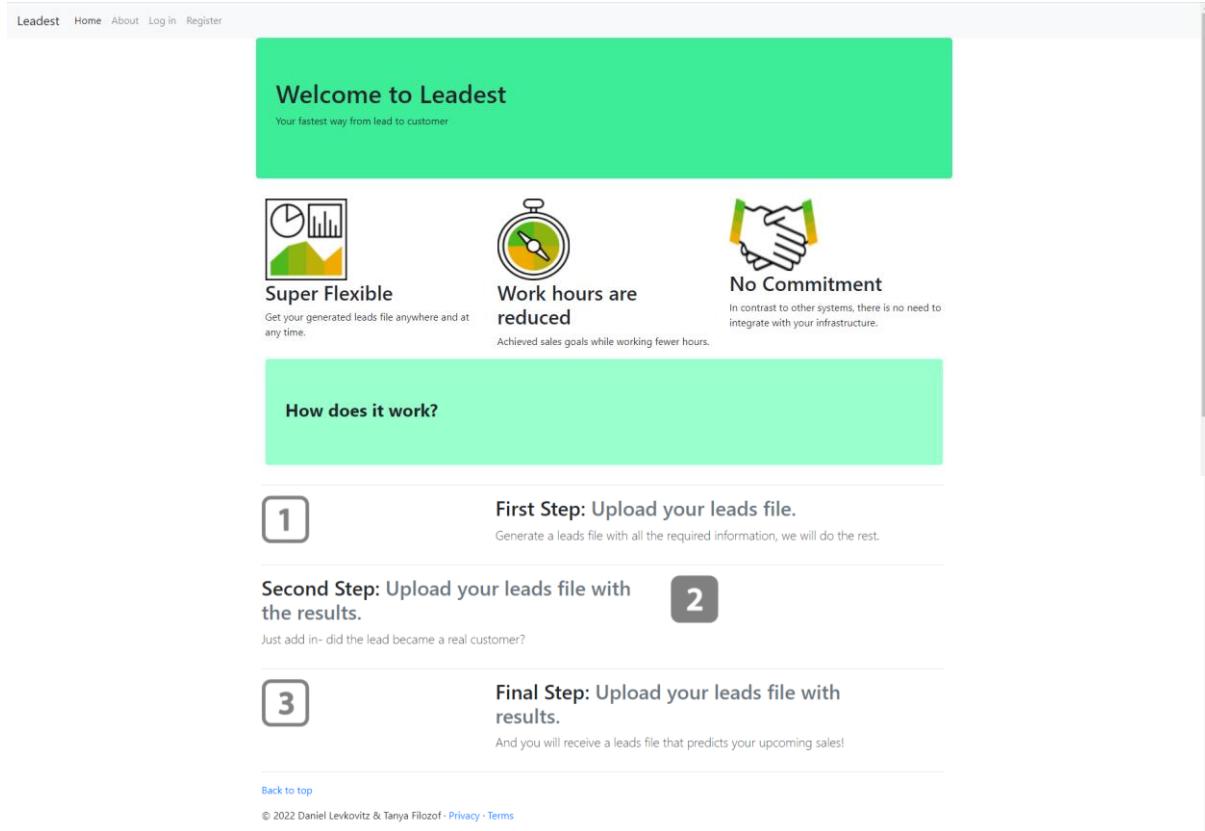
טכניקת חיפוש הרשת יוצרת מספר מודלים שונים של אותו אלגוריתם, בכל פעם עם היפר פרמטרים שונים, ומאמנת את המודל.

בסופו של דבר השיטה הזאת תחזיר את המודל בעל היפר פרמטרים שהצליחו לחזות עם אחוז הדיקט הגדול ביותר.

10 תוצר הפרויקט: מערכות מידע

10.1 מסכום של אתר האינטרנט

הדף מחולקים ומוסדרים לפי תרשימים עז המסכים [8.7](#).
10.1.1 דף הנחיתה של האתר לפני הרשמה והתחברות

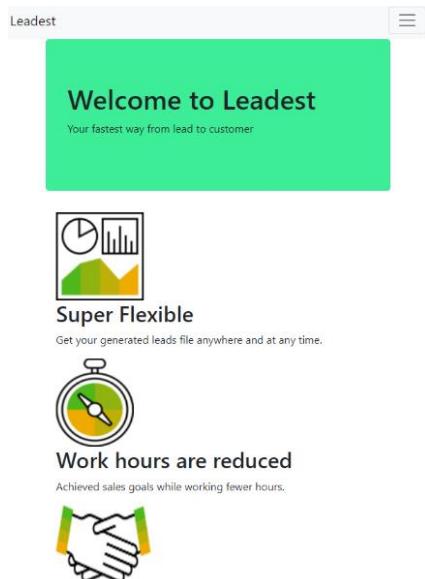


The screenshot shows the Leadest landing page. At the top, there's a navigation bar with links for Leadest, Home, About, Log in, and Register. The main content area has a green header with the text "Welcome to Leadest" and "Your fastest way from lead to customer". Below this, there are three sections: "Super Flexible" (with an icon of a chart and a bar graph), "Work hours are reduced" (with an icon of a stopwatch and a compass), and "No Commitment" (with an icon of two hands shaking). Each section has a brief description. Below these sections, there's a green box with the text "How does it work?". Underneath, there's a numbered list of steps:

- 1** First Step: Upload your leads file.
Generate a leads file with all the required information, we will do the rest.
- 2** Second Step: Upload your leads file with the results.
Just add in- did the lead became a real customer?
- 3** Final Step: Upload your leads file with results.
And you will receive a leads file that predicts your upcoming sales!

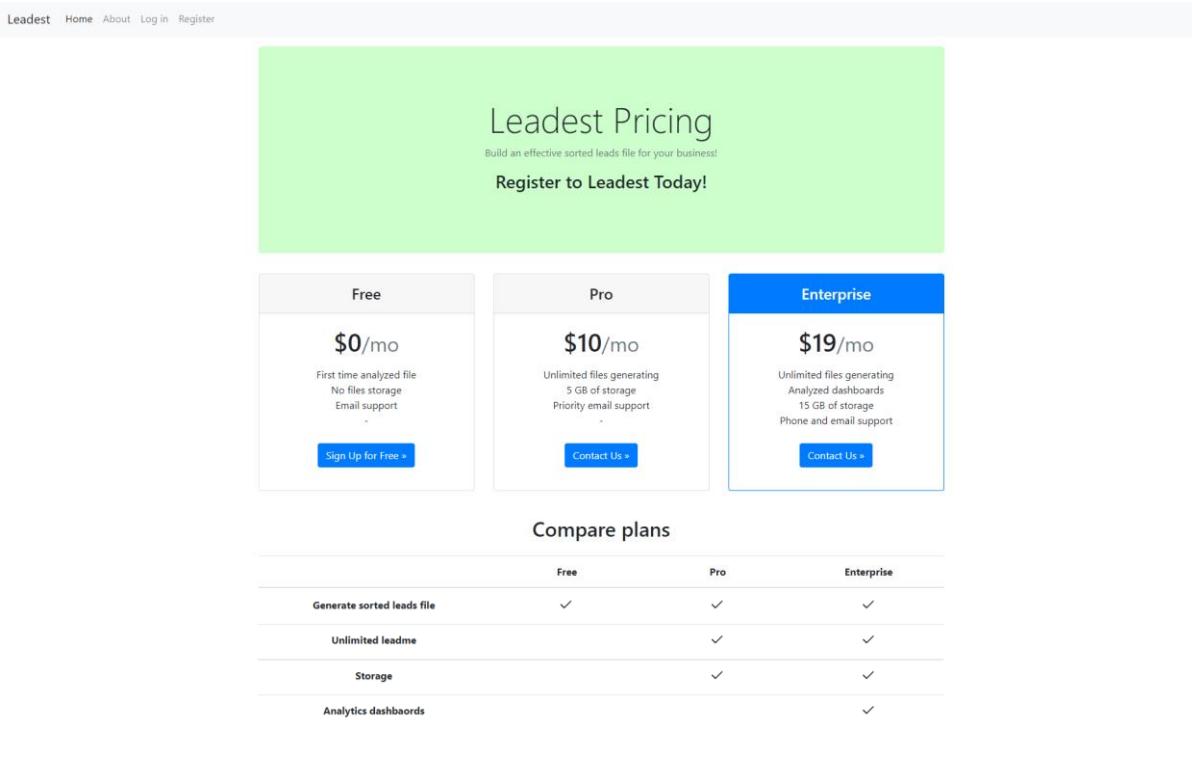
At the bottom of the page, there are links for "Back to top", "© 2022 Daniel Levkowitz & Tanya Filozof - [Privacy](#) · [Terms](#)".

דף הנחיתה של האתר מכיל מידע שיווקי והסביר על הعلاאת הקובץ לטובות מيون וקטלוג ושלבי התהליך. על מנת לשמר על רלוונטיות, האתר מזזה התחברות של המריכת וה- Navbar מציג קומפוננטות שונות כאשר משתמש מחובר לחשבון שלו. בנוסף, האתר מותאם למצב בו המשתמש מקטין את המסר והופך להיות קטן יותר (בכל עמוד):



The screenshot shows the Leadest landing page on a mobile device. The layout is compressed, with the "Welcome to Leadest" header and the three main features ("Super Flexible", "Work hours are reduced", and "No Commitment") appearing in a smaller, more compact form. The overall design is responsive, designed for mobile devices.

10.1.2 דף הסברים ומחירים של השירות



The screenshot shows the Leadest Pricing page. At the top, there's a green header with the text "Leadest Pricing" and a subtext "Build an effective sorted leads file for your business!". Below this is a button "Register to Leadest Today!". The main content area displays three pricing plans:

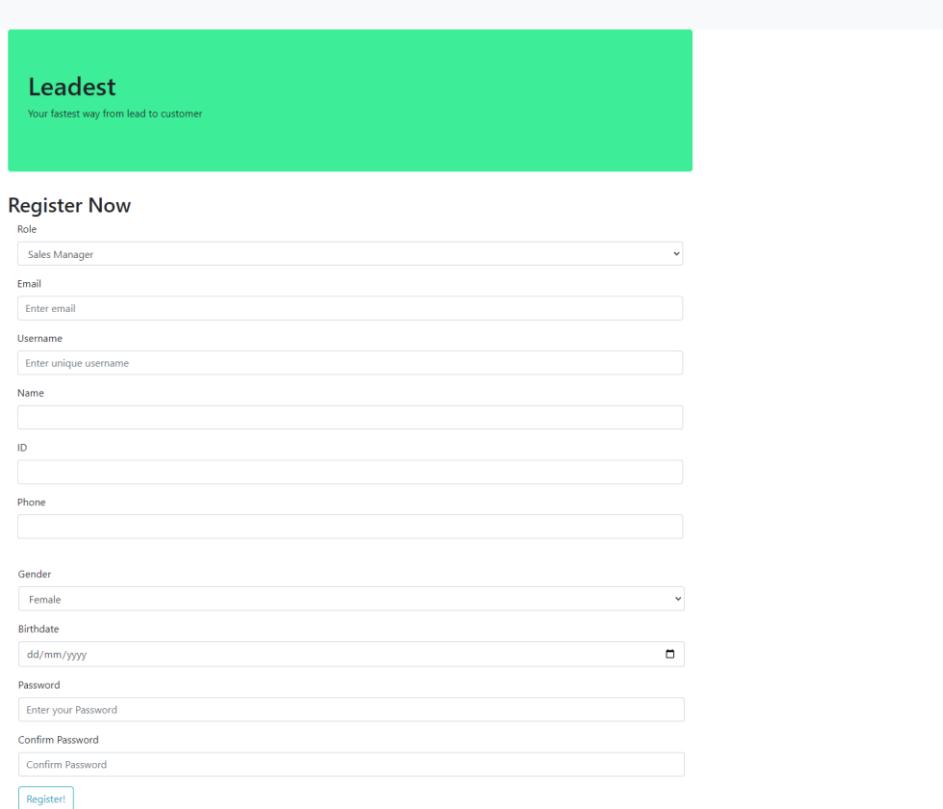
Free	Pro	Enterprise
\$0/mo	\$10/mo	\$19/mo
First time analyzed file No files storage Email support	Unlimited files generating 5 GB of storage Priority email support	Unlimited files generating Analyzed dashboards 15 GB of storage Phone and email support
Sign Up for Free »	Contact Us »	Contact Us »

Below the plans is a section titled "Compare plans" with a table:

	Free	Pro	Enterprise
Generate sorted leads file	✓	✓	✓
Unlimited leadme		✓	✓
Storage		✓	✓
Analytics dashbaords			✓

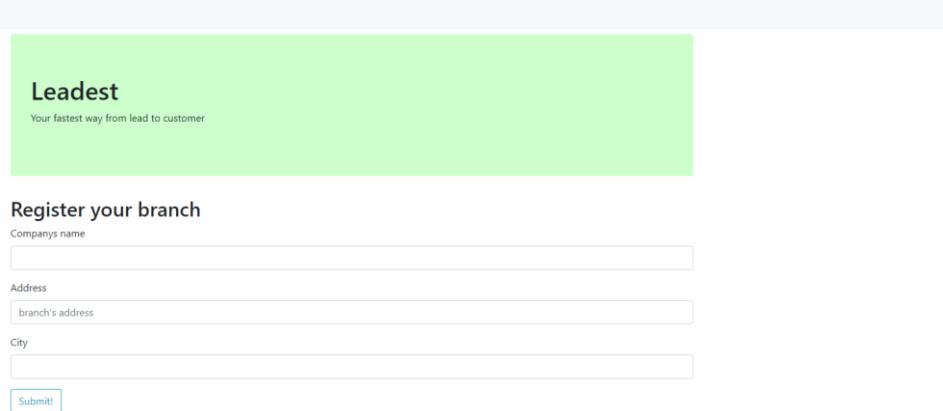
At the bottom of the page, there are links for "About", "Team", "Privacy", and "Terms".

דף המחירים נועד להסביר ללקוח חדש את המחירים של השירות והשוואה בין סוגי המנויים השונים. הלחיצנים הינם אינטראקטיביים והלקוח בעת לחיצה על כפתור על ההרשמה בחינם מועבר לעמוד ההרשמה וכפתור ה-[Contact Us](#) מפנה את הלקוח לשיחת מייל לדואר האלקטרוני שנוצר עבור הפרויקט.

10.1.3 הרשמה לאתר**10.1.3.1 הרשמה פרטיים אישיים**

The screenshot shows the 'Register Now' page of the Leadest application. At the top, there's a navigation bar with links for Leadest, Home, About, Log in, and Register. Below the navigation is a green header bar with the 'Leadest' logo and the tagline 'Your fastest way from lead to customer'. The main content area is titled 'Register Now'. It contains several input fields: 'Role' (set to 'Sales Manager'), 'Email' (placeholder 'Enter email'), 'Username' (placeholder 'Enter unique username'), 'Name' (placeholder 'Enter name'), 'ID' (placeholder 'Enter ID'), 'Phone' (placeholder 'Enter phone number'), 'Gender' (set to 'Female'), 'Birthdate' (placeholder 'dd/mm/yyyy'), 'Password' (placeholder 'Enter your Password'), and 'Confirm Password' (placeholder 'Confirm Password'). At the bottom of the form is a blue 'Register!' button.

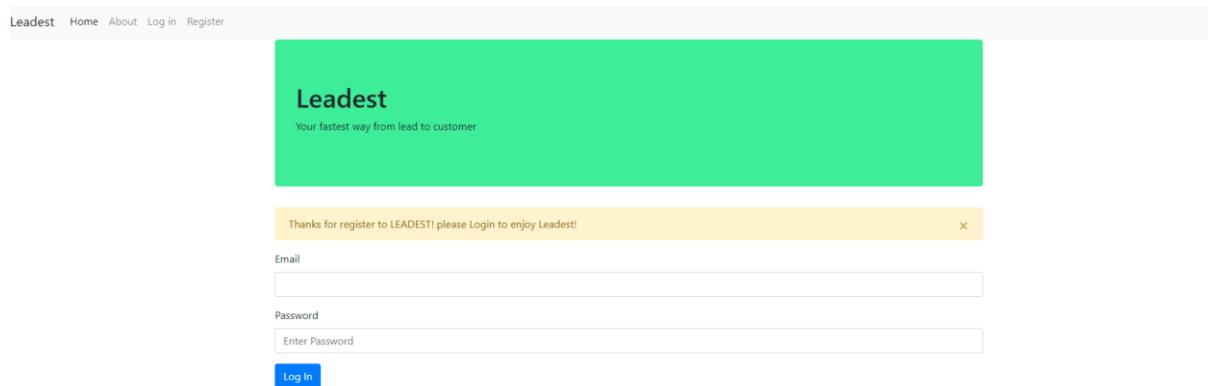
דף ההרשמה לאתר מכיל את כל השדות שהלוקהן צריך למלא בצדיו להירשם למערכת. במידה והוא כותב באחד השדות קלט לא תקין או שערק זה קיים כבר במערכת, המערכת לא מאפשרת להלוקהן להירשם ומופיעה הערה בהתאם. בעת סיום תקין ולחיצה על כפתור ה-Register, המערכת יוצרת אובייקט חדש במחלקה של Users, שומרת את נתוני המשתמש בטבלה של Users ב-DB ומעבירה על ידי פיתון ושימוש בחבילה של Flask את מספר המשתמש הייחודי שנוצר.

10.1.3.2 דף הרשמה לסניף

The screenshot shows the 'Register your branch' page of the Leadest application. At the top, there's a navigation bar with links for Leadest, Home, About, Log in, and Register. Below the navigation is a green header bar with the 'Leadest' logo and the tagline 'Your fastest way from lead to customer'. The main content area is titled 'Register your branch'. It contains three input fields: 'Company name' (placeholder 'Enter company name'), 'Address' (placeholder 'branch's address'), and 'City' (placeholder 'Enter city'). At the bottom of the form is a blue 'Submit!' button.

דף ההרשמה לסניף מועבר מסטר המستخدم, כאשר הוא צריך למלא את פרטי הסניף אליו הוא משתייך - שם החברה, כתובת ועיר. בעת לחיצת על כפתור ה-Submit של אותו ה-Form, המערכת יוצרת אובייקט מסוג Branch ומוסיפה את מספר המשתמש, בו זמנית המערכת שומרת את הנתונים של הסניף בDB ומעניקה לסניף מספר ייחודי.

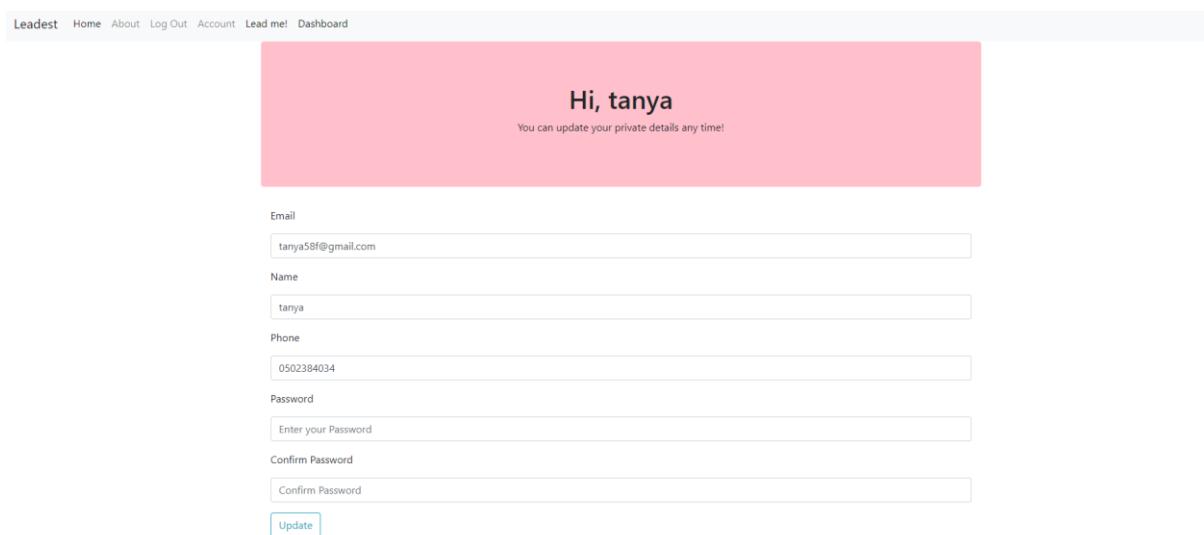
10.1.4 התחברות לאתר



The screenshot shows the Leadest login page. At the top, there's a navigation bar with links for Leadest, Home, About, Log in, and Register. Below the navigation is a green header bar with the Leadest logo and the tagline "Your fastest way from lead to customer". A yellow banner at the top of the main content area says "Thanks for register to LEADEST! please Login to enjoy Leadest!" with a close button (X). The main form has fields for Email (tanya50f@gmail.com) and Password (0502384034), followed by a blue "Log In" button.

לאחר ההרשמה, תופיע הודעה שמודה על ההרשמה ובקשת להתחבר. במידה והלkoח לא נרשם אלא רק נכנס לעמוד ה-Login, הודעה זו לא תופיע. בעת הת לחברות למערכת יופיעו הדפים הבאים:

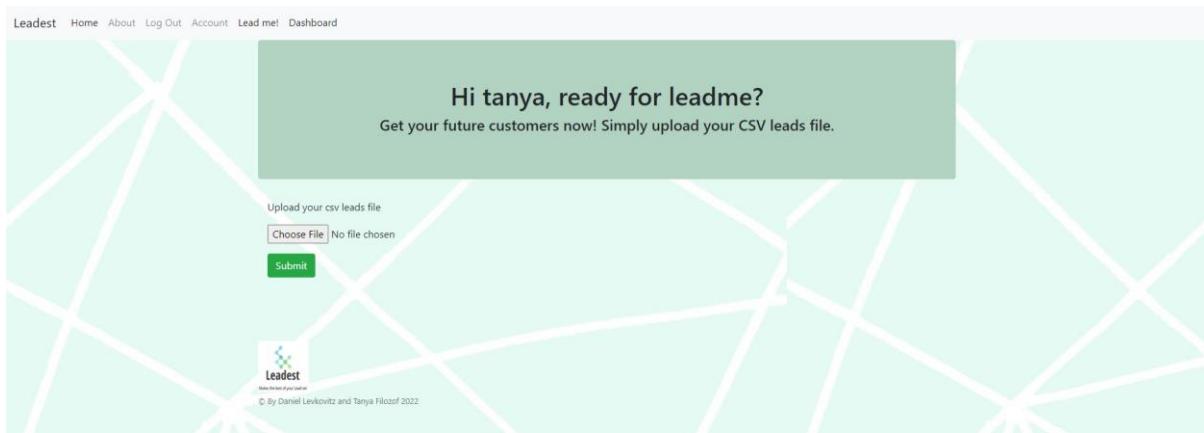
10.1.5 עריכת פרטי משתמש



The screenshot shows the Leadest account update page. At the top, there's a navigation bar with links for Leadest, Home, About, Log Out, Account, Lead me!, and Dashboard. Below the navigation is a pink header bar with the text "Hi, tanya" and the subtext "You can update your private details any time!". The main form contains fields for Email (tanya50f@gmail.com), Name (tanya), Phone (0502384034), Password (Enter your Password), and Confirm Password (Confirm Password). There is also an "Update" button.

הלkoח יכול לשנות בכל זמן נתון את פרטי המשתמש שלו: סיסמה, מייל, שם ומספר טלפון. השימוש בפינון מאפשר שימוש בחבילות של Flask ושימוש ב-Forms, ניתן להציג בטופס את הנתונים של הלkoח ולכן בעמוד ה-Account הלkoח יראה את הנתונים שמופיעים היום המערכת יוכל לעדכן את הנתונים בatabase מיד.

10.1.6 דף העלאת קובץ לידיים

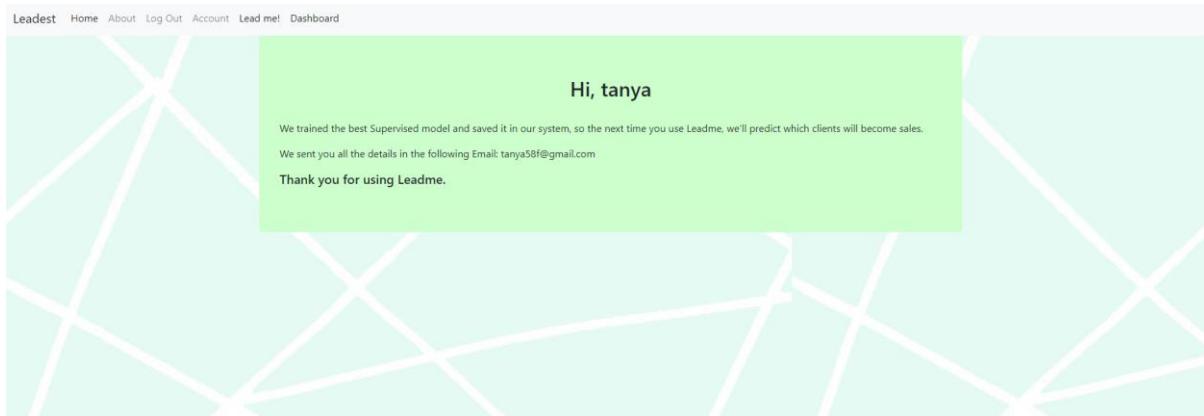


המערכת מזהה את הקובץ שהועלה ותציג דפים בסיום ההרצאה על פי הקובץ שהועלה:

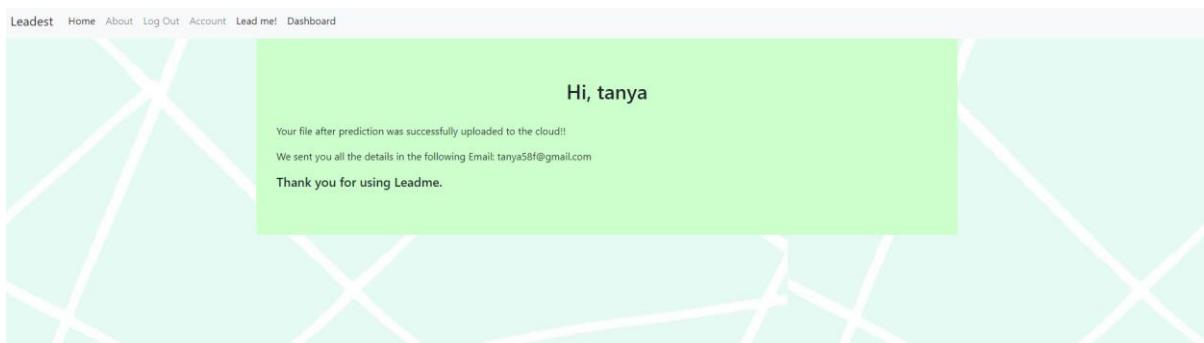
10.1.6.1 הupload קובץ ראשון



10.1.6.2 הupload קובץ שני



10.1.6.3 הupload קובץ שלישי



Dashboard מס' 10.1.7

Leadest Home About Log Out Account Lead me! Dashboard

Hi, tanya

Please upload a file you want to analyze its features

No file chosen



במרכז ה-**Dashboard** תופיע אפשרות להעלות קבצים והטסר יעבור לדף בהתאם לקובץ שהועלה.
עבור קובץ ללא תוצאות Dashboard 10.1.7.1

Leadest Home About Log Out Account Lead me! Dashboard

Hi tanya,

Are you ready to learn more about your data?

Lead analysis before sales attempts

The charts can be viewed by clicking

[Creation time vs. Is Business](#)

[Platform vs. Is Business](#)

[Car year vs. Year of birth](#)

[Upload another file](#)


Leadest
Makes the best of your Lead set
© By Daniel Levkowitz and Tanya Filozof 2022

המערכת מזיהה אם יש נתון אם הליד נמכר או לא, במידה והוא בעלי תוצאות יعلاה הדף זהה.

גרף של זמן יצירת הליד כתלות באם מדובר בליך עסק' או לא

10.1.7.2



דף זה הינו אחד עבור כל הגрафים, אך הפקנציונליות של הגרף משתנה בהתאם לграф הנבחר.
בgraf זה ניתן לראות כי ל��וחות עסק'ים פונים בשעות העבודה לעומת לקוחות פרטיים המשאירים פרטיים בשעות שלא מוגדרות כשעות עבודה.

גרף של פלטפורמה לעומת אם הליך עסק' או לא

10.1.7.3

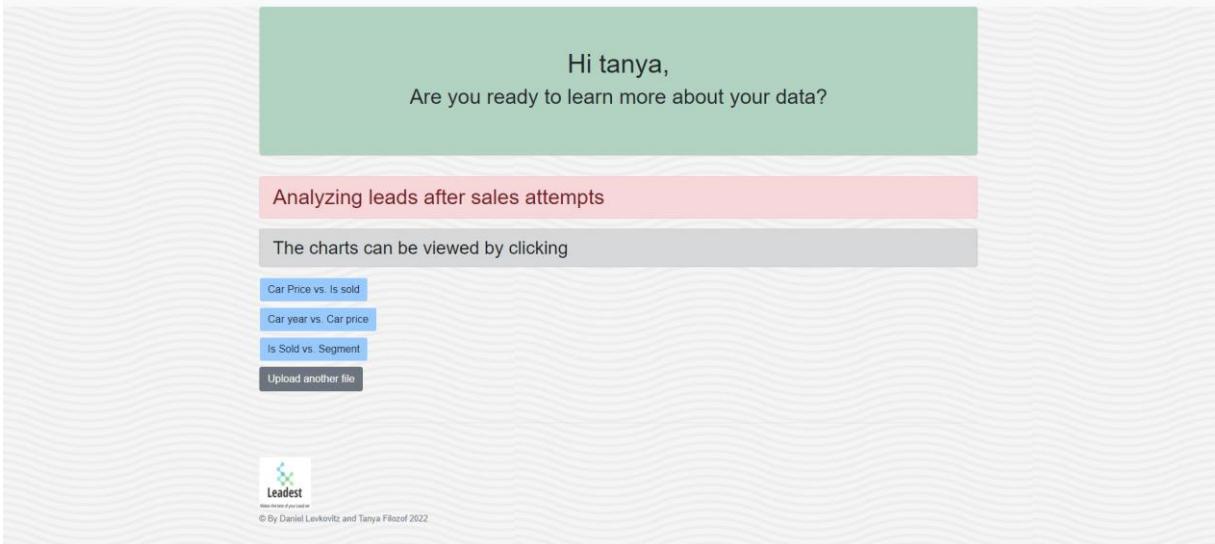


בgraf זה על פי קובץ הנתונים שהועלה, ניתן לראות כי ל��וחות עסק'ים פונים יותר דרך האתר החברה או דרך גוגל, בזמן של לקוחות פרטיים פונים על ידי שימוש בפלטפורמות חברותיות (ככל הנראה מהפונייה של פרסום החברה).

10.1.7.4 גרפ המתאר את שנת הלידה לעומת שנת ייצור הרכב ובנוסף התפלגות פלטפורמת הליד



10.1.7.5 עבור קובץ עם תוצאות Dashboard



10.1.7.6 גרפ שומרה תלות בין מחיר הרכב לבין אם נמכר או לא ומראה אם הליד הוא עסק או לא



Leadest
Makes the best of your Lead set
© By Daniel Levkovitz and Tanya Filozof 2022

הגרף ממחיש את השערתנו ההתחלתיות, שלידים עסקים הופכים למכירה ממשית בסיסי גובה יותר.

10.1.7.7 גרפ של שנת הרכב לעומת מחיר הרכב

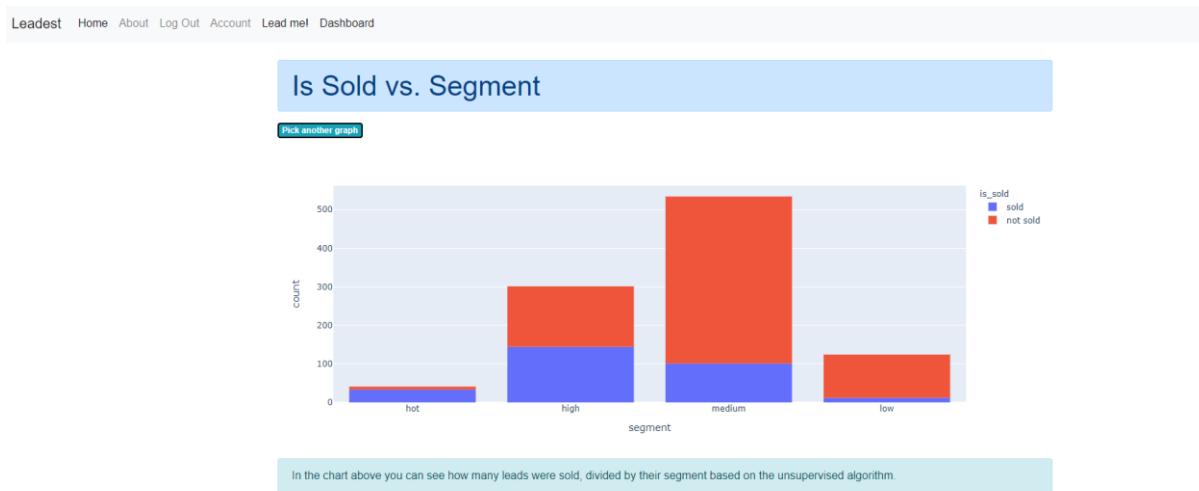


Leadest
Makes the best of your Lead set
© By Daniel Levkovitz and Tanya Filozof 2022

ניתן לראות בגרף זה כי רכבי יוקרה אשר מחירם גבוהה יותר, נמכרים פחות לעומת רכבים זולים עם שנותן מאוחר יחסית.

גרף שמציג את החלוקה של הקבוצה של הלקוחות לצד לעומת אם נמכר או לא

10.1.7.8



Leadest
Makes the best of your Lead set
© By Daniel Levkovitz and Tanya Filozof 2022

10.2 תוכרים מען ב-Google Cloud Platform

הוות ונעשה שימוש בAPI'S ובסלטפורמת הענן של GOOGLE, נוסיף את התוכרים של דוגמאות מהענן אליו המשתמש של המוצר מחובר.

Final project leads Bucket 10.2.1

Google Cloud Platform Select a project ▾

Cloud Storage Bucket details

final_project_leads

OBJECTS CONFIGURATION PERMISSION PROTECTION LIFECYCLE

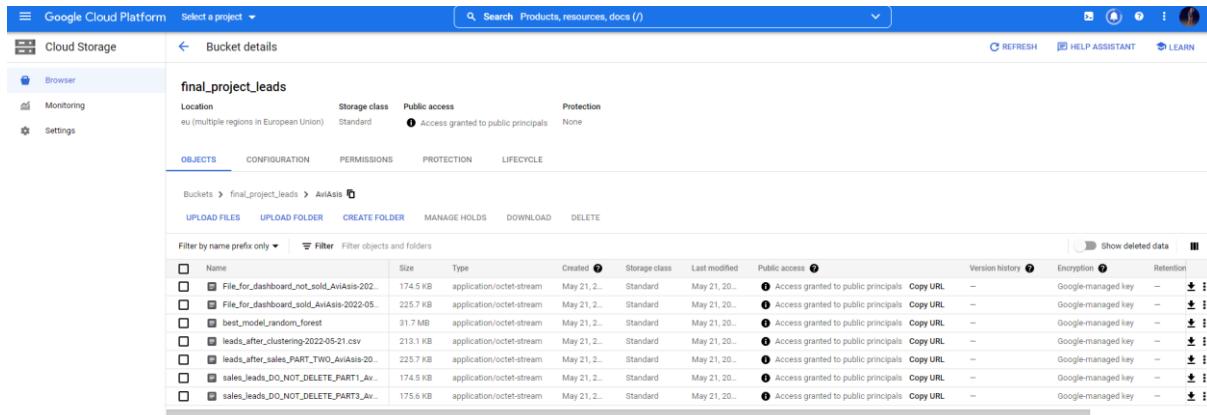
Buckets > final_project_leads

UPLOAD FILES UPLOAD FOLDER CREATE FOLDER MANAGE HOLDS DOWNLOAD DELETE

Name	Size	Type	Created	Storage class	Last modified	Public access	Version history	Encryption	Retention expiry date	Holds
AviAsil/	—	Folder	—	—	—	—	—	—	—	—
Car_Leads.csv	157.6 KB	text/csv	11 May 2022, 22:15:25	Standard	11 May 2022, 22:15:25	Value hidden	Copy URL	Google-managed key	—	None
Companies Data.csv	137.2 KB	text/csv	6 May 2022, 15:18:55	Standard	6 May 2022, 15:18:55	Value hidden	Copy URL	Google-managed key	—	None
danielle/	—	Folder	—	—	—	—	—	—	—	—
daniellewko/	—	Folder	—	—	—	—	—	—	—	—
tanya/	—	Folder	—	—	—	—	—	—	—	—
vehicles4.csv	17.8 MB	text/csv	6 May 2022, 15:19:05	Standard	6 May 2022, 15:19:05	Value hidden	Copy URL	Google-managed key	—	None

לאחר העלאת הקובץ לאתר, האתר יוצר "דלי" (כמו תיוקיה בטרמינולוגיה של מחשב ענן) ייעודי עבור המשתמש שאליו יעלטו כל הקבצים, בנוסף, האלגוריתם משתמש בקבצים שנמצאים מחוץ לתיקיות של המשתמשים לטובות חיזוי הקובץ.

10.2.2 תקיה של משתמש



The screenshot shows the Google Cloud Platform Cloud Storage interface. The bucket 'final_project_leads' is selected. The table below lists several objects:

Name	Type	Created	Storage class	Last modified	Public access	Version history	Encryption	Retention
File_for_dashboard_not_sold_AviAisis-202...	application/octet-stream	May 21, 20...	Standard	May 21, 20...	Access granted to public principals	Copy URL	Google-managed key	
File_for_dashboard_sold_AviAisis-2022-05...	application/octet-stream	May 21, 20...	Standard	May 21, 20...	Access granted to public principals	Copy URL	Google-managed key	
best_model_random_forest	application/octet-stream	May 21, 20...	Standard	May 21, 20...	Access granted to public principals	Copy URL	Google-managed key	
leads_after_clustering-2022-05-21.csv	application/octet-stream	May 21, 20...	Standard	May 21, 20...	Access granted to public principals	Copy URL	Google-managed key	
leads_after_sales_PART_TWO_AviAisis-202...	application/octet-stream	May 21, 20...	Standard	May 21, 20...	Access granted to public principals	Copy URL	Google-managed key	
sales_leads_DO_NOT_DELETE_PART1_Avi...	application/octet-stream	May 21, 20...	Standard	May 21, 20...	Access granted to public principals	Copy URL	Google-managed key	
sales_leads_DO_NOT_DELETE_PART3_Av...	application/octet-stream	May 21, 20...	Standard	May 21, 20...	Access granted to public principals	Copy URL	Google-managed key	

בתיקיה ניתן לראות את כל הקבצים שהועלו בשימוש ב-API הקיימים את 7 הקבצים: הקבצים המשמשים את הגرافים בעמוד Dashboard, שם המודל שנבחר להיות הטוב ביותר בשלב השני, את הקבצים שהלכו הعلاה בעבר שלב ראשון ושלישי ואת הקבצים שהמערכת העלתה לאחר למידה מונחת ולמידה בלתי מונחתית.

10.3 פלטי אימיל

10.3.1 לאחר הulat קובץ ראשוני

Leadest - Clustered CSV File ➜ Inbox

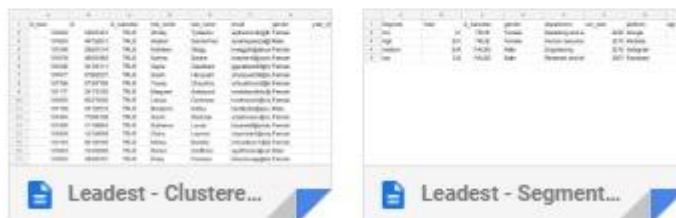
 leadest.leadme@gmail.com
to me ▾

Thanks for using Leadme!
your Clustered CSV is attached to this email.

Additionally, we sent you a CSV summary of what each Lead type means!

Greeting,
Leadest

2 Attachments



המשתמש לאחר הulat הקובץ הראשוני ולאחר שימוש באלגוריתם של למידה בלתי מוחנית, מקבל מייל עם תוכנות ההרצתה עם הסבר.

10.3.2 הסיכום של הלידים לאחר החלק הראשוני

profit	Market Cat	time_cata	desirable_car	car_price	age	platform	car_year	department	gender	is_buisnes	Total	Segment
22527.63	401741	Morning	607	107812.6	40	Google	2020	Marketing	Female	TRUE	41	hot
633.6684	14703.54	Noon	548	14744.32	48	Website	2015	Human res	Female	TRUE	301	high
421.5114	9222.086	Night	233.5	12991.3	61	Instagram	2015	Engineering	Male	FALSE	534	medium
-214.651	13331.84	Evening	242	7188.07	54.5	Facebook	2007	Research	Male	FALSE	124	low

המשתמש מקבל מייל לאחר ההרצתה הראשונה, סיכום של החלוקה ומשמעותם בקובץ CSV.
בקובץ מופיעים כל האשכולות וכל אשכול מכיל את הערך השכיח בעמודות הקטגוריאליות והממוצע בקטגוריות הנומריות - על מנת שהמשתמש יבין את התפלגות האשכולות.

10.3.3 מיל לאחר בחירת המודל הטוב ביותר

Leadest - Model was trained  Inbox x

 **leadest.leadme@gmail.com**
to me ▾

Thanks for using Leadme!

We learned the best model which is random_forest and saved it in our cloud.

The accuracy of the learning model was 0.832

Next time you will use Leadme we will predict which lead will become a sell!

Greetings
Leadest

לאחר העלאת הקובץ השני, המשתמש מקבל מיל על המודל הטוב ביותר שנבחר עבור הקובץ הכלול את תוצאת הדיק שמודל קיבל.

10.3.4 מיל לאחר החלק השלישי

Leadest - Predicted CSV File  Inbox x

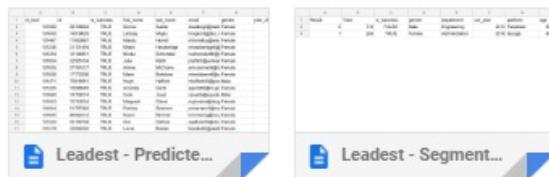
 **leadest.leadme@gmail.com**
to me ▾

Thanks for using Leadme!

We predicted for you which lead will become a sell!

Greetings
Leadest

2 Attachments



לאחר החלק האחרון המשתמש מקבל מיל המכיל את קובץ הלידים החזוי והסיכום.

10.3.5 סיכום לאחר החלק השלישי

profit	Market Ca	desirable	car_price	age	platform	car_year	departmen	gender	is_buisnes	Total	Result
538.1243	12359.71	252	12688.99	58.81921	Facebook	2014	Engineerin	Male	FALSE	708	0
3191.252	64124.01	545	26380.92	49.20548	Website	2015	Human res	Female	TRUE	292	1

המשתמש מקבל את קובץ הסיכום למייל הכלול את החלוקת לשתי קבוצות: יימכר או לא על פי מודל החזוי שנבחר. הקובץ מכיל את הערך השכיח בעמודות הקטגוריאליות והממוצע בקטגוריות הנקודות - על מנת שהמשתמש יוכל את התפלגות שתי הקבוצות.

11 תובנות כריית המידע והמודלים שפותחו עבורה

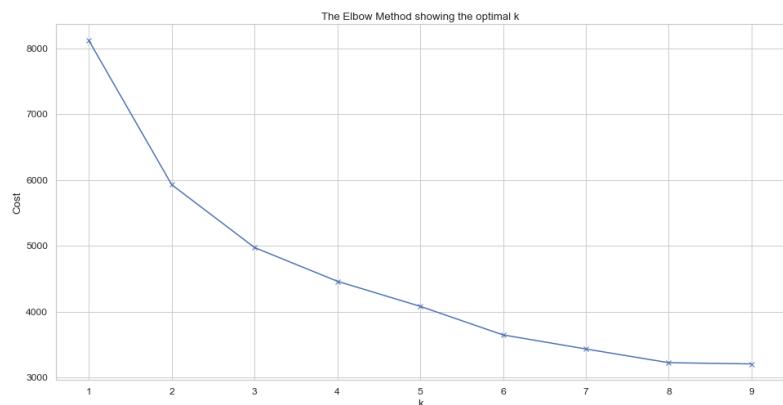
בחלק זה, נפרט אודות תובנות כריית המידע ואילו מודלים פותחו בשבייל לבצע את המערכת. חלק זה יתמקד בעיקר איך מtabצע חישוב האלגוריתמים במערכת, מה היו התוצאות של כריית המידע מבחינה סטטיסטיות, ויום מבחינת הקוד. נציין שפירטנו על הנитוחים הסטטיסטיים המקדימים לביצוע האלגוריתמים בסעיף [תיקוף ובדיקות בסעיף 6.2](#).

- פירטנו על תהליך יבוא הנתונים, ETL וה-Pre-processing בסעיף [7.1](#).
 במקרה זה נתיחס לתובנות של כריית המידע בשלושת החלקים של המערכת:
 I. החלק הראשון בלמידה הבלתי מונחית כאשר האלגוריתם עדין לא יודע איזה ליד יփוך למכירה והוא משתמש באלגוריתם K – Prototypes
 II. החלק השני בלמידה המונחית כאשר אנשי המכירות עידכו את קובץ הלידים על פי המכירות ונבחר עבורם האלגוריתם הטוב ביותר.
 III. החלק השלישי כאשר נבחר מודל הלמידה המונחית הטוב ביותר, אנחנו חוזים אילו לידים יפכו למכירה.

11.1 החלק הראשון – תובנות כריית המידע של אלגוריתם K – Prototypes

כפי שציינו בהקדמה, לאחר יבוא הנתונים של: קובץ הלידים המקורי, מאגר הרכבים, ומאגר החברות העסקיות, לאחר ניקוי הנתונים, מיזוגם וביצוע נרמול על העמודות הנומריות, לפני הכנסתםkeklat לאלגוריתם, היינו צריכים להחליט מהו מספר הקבוצות האופטימלי שנחלק את הלידים שלנו כדי שנוכל לאפיין כל קבוצה, ולקטול אותה. בשבייל להחליט על מספר הקבוצות השימוש ביטה ממדעי הנתונים שנקראת שיטת המרפק.

11.1.1 שיטת המרפק לבחירת ה-K האופטימלי



גרף 11.1.1. שם שיטת המרפק

מידע על שיטת המרפק בסקרת הספרות [3.6.7](#).

בשיטת המרפק אנו ננסח לבחור K מרכזים שונים, בכל איטרציה רצים על K אחר (בדוגמה כאן מ-1 עד 9), ובכך מחשבים כמה הנתונים, שבמקרה שלנו הם הלידים, רוחקים מהמרכז שלהם ככל שאנו מגדילים את מספר הקבוצות.

ציר-X אצלנו זה ה-K כולל לכמה מרכזים שונים אנחנו מחלקים את הדטה, וציר-Y זה השגיאה הכוללת של הלידים מ-K המרכזים.

מכיוון שאנחנו משתמשים באלגוריתם K-Prototypes השגיאה של העמודות הקטגוריאליות מחושבת על ידי ממד האי דמיון, והשגיאה של העמודות הנומריות מחושבות על ידי מרחק האוקלידי.

ניתן לראות שבגלל שכרגע הנתונים שלנו מאוד מפוזרים, ככל ש-K עולה כך גם השגיאה יורדת.

עובדה זו אינטואיטיבית כי אם נחלק את הלידים שלנו להרבה קבוצות המרחק בין הלידים לכל קבוצה יהיה נמוך יותר, אך לעומת זאת, לחלק למספרים מומס קבוצות לא יהיה אפקטיבי כי אנחנו נרצה למצאו דמיון בין כל קבוצה ובכך לסתווג את הלידים שלנו.

בסוף דבר בהשראת גרפ זה, בחרנו ב- $4 = K$ כי ניתן לשים לב שב $1 = K$ השגיאה הייתה מעלה-8000, כאשר אנחנו ב- $4 = K$ השגיאה עומדת על 4400, והשגיאה המינימלית שמתאפשרת ב- $9 = K$ היא 3200.

מכיוון שאין הבדל ממשוני שאנחנו עוברים את $4 = K$, בחרנו במספר הקבוצות הזה, מה שיתמוך גם בהנחה שביצענו בהתחלה שהיא לנו 4 קבוצות של ילדים.

11.1.2 שימוש באלגוריתם K – Prototypes

הסבר באופן כללי על אלגוריתם Prototypes – K אפשר לקרוא בסקירת הספרות [3.5.6](#)

لتיאור מלא של אופן חישוב אלגוריתם בסעיף [9.2](#).

בסעיף זה נתמקד איך יישמו את המודל בשפת התכנות פיתון ומה ההיפר פרמטרים שבחרנו למודל.

Using K-Prototypes

```
In [63]: kproto = KPrototypes(n_jobs = -1, n_clusters = 4, random_state=6, init = 'Cao', verbose=1,max_iter=100,n_init=10)
clusters = kproto.fit_predict(df_leads_for_analysis, categorical=categorical_columns)

Initialization method and algorithm are deterministic. Setting n_init to 1.
Best run was number 4
```

דבר ראשון נשים לב שאנחנו משתמשים בפונקציית API שנראית "KPrototypes" אשר מיובאת מהספרייה "kmodes", ספרייה חינמית בפייתון אשר מכילה אלגוריתמי אשכולות שונים. לאחר שהגדכנו את המודל, אנחנו צריכים להכנסו לו את ההיפר פרמטרים, שעל פיהם, האלגוריתם יידע כיצד לבצע את החישוב בכל איטרציה.

11.1.2.1 הסבר על כל היפר פרמטר בטבלה הנ"ל:

הסביר	היפר פרמטר
מספר המעבדים לשימוש אופן ביצוע וחישוב האלגוריתם. כאשר 1 - זה פרמטר שומר שבו כל המעבדים שנגשימים באותו רגע משמשים לחישוב האלגוריתם.	jobs_N
מספר האשכולות שנרצה לחלק	n_clusters
גרעין אקראי, מכיוון שבחרית האשכולות הראשיים הראשונים קורית בצורה רנדומלית, ובכל הרצת יכול להתקבל תוצאה שונה, הגרעין האקראי גורם לזה שככל פעם שנ裏ץ, אותו אשכולות ראשיים ראשוניים יבחרו ומשם האלגוריתם ימשיך בשאר האיטרציות.	random_state
סוג החישוב של פונקציית ההפסד, אנחנו בחרנו ב-Cao.	init
פרמטר בוליани, אם בוחרים ב-1, הוא יראה את אופי החישוב מאחוריה הקלעים וכמה פונקציית ההפסד יורדת בכל איטרציה.	verbose
מספר האיטרציות המקסימלי בריצה הנוכחית.	max_iter
מספר הפעמים שהאלגוריתם יופעל עם מרכזים שונים (כלומר בכל פעם הוא מתחילה באופן רנדומלי עם בחירת מרכזים שונים) בסופה של דבר מתקיים המרכזים ההתחלתיים עם ערך בפונקציית ההפסד הנמוכה ביותר.	n_init

טבלה 11.1.2.1. הסבר היפר פרמטרים של K – Prototypes

לאחר שהגדכנו את ההיפר פרמטרים כלומר את סט החוקים שעל פייהם יפעל האלגוריתם, אנחנו משתמשים בפונקציית Predict, כלומר על הדאטה סט של הלידים, תחשב את המידע הזה ועל פיו תחשב לי את המרכזים.

נשים לב שם שnoch בשימוש בפונקציה זו מתוך הספרייה kmodes, שמכיוון שהאלגוריתם יודע להתמודד עם ערכים נומריים וקטגוריאליים, רק צריך לציין לו קלקט מה האינדקס של העמודות הקטגוריאליות, והוא יידע בעמודות אלה לחשב את ממד האי דמיון כפונקציית ההפסד, ובעמודות הנומריות הוא יחשב את המרחק האוקלידי.

11.1.3 המרכזים שנבחנו

Inspect our Centroids

```
In [78]: print(kproto.cluster_centroids_)

[[ '0.23612501266667824' '0.30333350025723604' '-0.18020339920307796'
  '-0.5581484864008706' '-0.16441315631722894' '-0.1555630795261142'
  'False' 'Male' 'Instagram' 'Engineering' 'Night']
 [ '1.4970076559513423' '-0.7538621859815131' '4.443271996306414'
  '1.3632411515358707' '3.913260657749725' '3.189756479908152' 'True'
  'Female' 'Google' 'Administration' 'Morning']
 [ '-1.9473597127276376' '-0.03462104071915656' '-0.4631685247481902'
  '-0.5552192048528864' '-0.281759020081989' '-0.1205369370981362'
  'False' 'Male' 'Facebook' 'Research and development' 'Evening']
 [ '0.17941705554887477' '0.42119179555795955' '-0.09472637742639285'
  '1.033241149258644' '-0.12527888041253235' '-0.10884634886751013'
  'True' 'Female' 'Website' 'Human resources' 'Noon']]
```

לאחר שהאלגוריתם סיים את החישוב, הוא בחר את המרכזים שיופיעו כל קבוצה, נשים לב שהמספרים לא אינטואיטיביים מכיוון שנחנכו הפעלו את האלגוריתם על הדאטה סט לאחר תהליך Pre-Processing שבו נormalנו את הנתונים (לכן הנתונים המספריים שנחנכו רואים הם ציון תקן).

אך לאחר שנשים כל אשכול ליד המקורי (כלומר על הדטה הנו MRI לפני שנרמלנו את הנתונים), נוכל לבדוק מה מופיע כל אשכול.

בסקירה קצרה של האשכול הראשון (כלומר האיבר הראשון במערך הדוח מדי), הערכים הקטגוריאליים השכיחים שמאפיינים את אשכול זה, הם לרוב לקוחות לא עסק'ים, גברים, שהגיעו מפלטפורמת אתר הנחיתה Instagram, עובדים במחלקות הנדסה והשairoו פרטימ בשעות הלילה.

11.1.4 שייר האשכולות לדטה פריים וסיכון

בשלב הראשון נרצה לחת את מערך האשכולות שהוא בעצם מערך חד ממד' שמכיל מספרים מ-0 עד 3 (בגל שמספר האשכולות שבחרנו הוא 4), ולשייר את תוכנות האשכולות לדטה פריים לפני שנרמלנו את הנתונים.

Analyizing our KPI's for the leads

```
In [82]: result_data=df_leads_app.copy()
result_data['lead_type']=clusters
```

יצרנו דטה פריים שהוא עותק של הדטה פריים המקורי לפני שביצענו נרמול נתונים, ויצרנו עמודה חדשה 'lead_type' כאשר העמודה זו תציג כל שורה בדטה שהיא ליד, לאיזה אשכול היא נבחרה להשתיר על ידי האלגוריתם.

Divide the clusters by their results to Hot, High, Medium, and Low

```
In [83]: result_data['Segment'] = result_data['lead_type'].map({0:'First', 1:'Second', 2:'Third',3:'Forth'})
# Order the cluster
result_data['Segment'] = result_data['Segment'].astype('category')
result_data['Segment'] = result_data['Segment'].cat.reorder_categories(['First','Second','Third','Forth'])
```

לאחר מכן, נרצה להמיר את שמות האשכולות מסוג דטה נומי, לשוג דטה קטגוריאלי כדי שהיא יהיה לנו נוח וקל יותר לנתח אותו.

לכן, הגדרנו עמודה חדשה שנקראת 'Segment' ששויה לעמודת ה-'lead_type' אך ביצענו שנייניו לשמות בסדר הבא: 0 – First , 1 – Second , 2 – Third , 3 – Fourth . הגדרנו שסוג העמודה תהיה מסוג "Category" , והגדירנו את הסדר שבו אנחנו רצימ שיויעו הקטגוריות (מ- First עד ל- Fourth).

כעת, שהאשכולות מסודרים בדטה פריים בצורה קטגוריאלית לפי סדר, ניתן להפעיל פונקציית הקבוצה (Aggregation) אשר נתונה בספרייה Pandas, ולקבץ את הנתונים לפי פונקציית ההקבוצה שנבחרה.

Creating a temp Dataframe to give a score for each attribute

```
In [82]: result_data.rename(columns = {'lead_type':'Total'}, inplace = True)
df_groupby_segment=result_data.groupby('Segment').agg(
{
    'Total':'count',
    'is_business': lambda x: x.value_counts().index[0],
    'gender': lambda x: x.value_counts().index[0],
    'department': lambda x: x.value_counts().index[0],
    'car_year': 'median',
    'platform': lambda x: x.value_counts().index[0],
    'age': 'mean',
    'car_price': 'mean',
    'desirable_rental_days': 'mean',
    'time_catagor': lambda x: x.value_counts().index[0],
    'Market Cap':'mean',
    'profit':'mean'
})
.reset_index()
df_groupby_segment
```

Segment	Total	is_business	gender	department	car_year	platform	age	car_price	desirable_rental_days	time_catagor	Market Cap	profi
First	834	False	Male	Engineering	2016.0	Instagram	80.992509	12991.303905	230.481311	Night	9222.088330	421.51142
Second	41	True	Female	Marketing and sales	2020.0	Google	43.634146	107812.555854	606.219512	Morning	401741.024380	22527.63414
Third	124	False	Male	Research and development	2007.0	Facebook	55.443548	7188.089758	231.024194	Evening	13331.835484	-214.65080
Forth	301	True	Female	Human resources	2016.0	Website	49.096346	14744.322558	541.681063	Noon	14703.538213	633.68843

לכן ייצרנו עמודת Total שמייצגת את כמות הלידים באותו אשלול, והשתמשנו בפונקציית `.Count` לעמודות הקטגוריאליות שהן: `is_buisness`, `time_catagor`, `platform`, `department`, `gender`, `car_price`, `age`, `desirable_rental_days`, `time_catagor`, `profit`, `Market Cap`, `car_year`, `car_price_score`, `platform_score`, `age_score`, `desirable_rental_days_score`, `time_catagor_score`, `profit_score`, `Market Cap_score`. השמשנו בשכיח, ככלומר הקטגוריה שהופיעה הכי הרבה פעמים תציג לי את המאפיין של אותו אשלול.

לעמודות הנומריות שהן `age`, `car_price`, `desirable_rental_days`, `car_year`, `Market Cap`, `profit`, `is_buisness`, `time_catagor`, `platform`, `department`, `gender`, השמשנו בממוצע בשבייל לייצג את אותו אשלול. ולעמודת `car_year`, גם השמשנו בחזין מכיוון שנתה הרכב הוא מספר שלם ולא משתנה רציף.

לבסוף, קיבלנו את הדאטה פריטים הבא שמסכם בצורה ברורה את תוצאות האלגוריתם ונוכל להשתמש בתוצאות אלה בשבייל להבין מהם לדיים "רותחים", לדיים "חמים", לדיים "בנייה" ולידים "קרים".

נzieין שאט הניתוח של הדאטה פריטים עצמו ושל הנתונים שהתקבלו סיכמנו בסעיף **תיקוף** ובדיקות והציגת הניתוחים הסטטיסטיים [6.2.1.1.6](#).

11.1.5 ניקוד התוצאות

לאחר שייצרנו את הדאטה פריטים המסכם את תוצאות האשלולות, נוכל להשתמש בטבלה זו בשבייל לצורך פונקציה שתחשב באופן אוטומטי מהו לדי "רותח", לדי "חם", לדי "בנייה" ולדי "קר".
לכן המטרה היא ליצור דאטה פריטים נוסף מהדאטה פריטים סיכום התוצאות, אשר באמצעות סיכומי הנתונים נוכל להשוות בין האשלולות ולקבוע לפי סדר מה הקטלוג שקיבל כל אשלול.

Scoring the results

```
In [159]: df_groupby_segment_temp=pd.DataFrame()
important_labels=['is_business','department','car_year','platform','age','car_price','desirable_rental_days','time_catagor','profit','Market Cap']
new_labels=[x+'_score' for x in important_labels]
df_groupby_segment_temp[important_labels]=df_groupby_segment[important_labels].apply(lambda x: 1 if x==True else 0)
for i,j in zip(important_labels,new_labels):
    if df_groupby_segment[i].dtypes==np.float64:
        if i=='age':
            df_groupby_segment_temp[j]=df_groupby_segment[i].rank(method='min',ascending=False)
        continue
        df_groupby_segment_temp[j]=df_groupby_segment[i].rank(method='max')
    else:
        if i=='is_business':
            df_groupby_segment_temp[j]=df_groupby_segment[i].apply(lambda x: 1 if x==True else 0)
        elif i=='department':
            df_groupby_segment_temp[j]=df_groupby_segment[i].apply(lambda x: 1 if x in['Administration','Marketing and sales','Human resources'] else 0)
        elif i=='platform':
            df_groupby_segment_temp[j]=df_groupby_segment[i].apply(lambda x: 1 if x in['Google','Website','Phone'] else 0)
        elif i=='time_catagor':
            df_groupby_segment_temp[j]=df_groupby_segment[i].apply(lambda x: 1 if x in['Morning','Noon','After Noon'] else 0)
```

בשביל לחשב זאת בצורה אוטומטית על ידי פונקציה, דבר ראשון ייצרנו עותק של הדאטה פריטים סיכום התוצאות, ייצרנו 2 רשימות שמקילות את שמות העמודות, כי אנו הולכים ליצור דאטה פריטים אשר יציג את הדירוג שקיבל כל אשלול, וצריך להכניס לדאטה פריטים שמות לעמודות.

כעת, נוח בולאה על רשימת העמודות, ובכל עמודה על כל הערכים.

אם הערכים הם נומריים: יתקבל ציון 4 עבור הערך הגבוה ביותר וציון 1 עבור הנמוך ביותר.

אם הערכים הם קטגוריאליים: בחלק זה נדרש להתייחס לכל עמודה בנפרד וכן נפריד בין סוג העמודות. העמודה `is_buisness` זו אשר מייצגת האם הלוקו העסקי, אז אם הוא עסק יתקבל הציון 1, אם לא יתקבל הציון 0.

בעמודת `department` אשר מייצגת את המחלקה שבה עובד הליד, אם הליד עובד במחלקות של אדמיניסטרציה, שיווק ומכריה, ומشاءבי אנוש, יתקבל הציון 1, אחרת, יתקבל הציון 0.

בעמודת `time_catagor` אשר מייצגת את הפרק זמן ביום שנוצר הליד, אם הליד נוצר בבוקר, צהרים או אחר הצהרים, יתקבל הציון 1, אחרת יתקבל הציון 0.

בעמודת 'platform', אשר מייצגת את אתר הנחיתה ממנה הגיע הליד, אם הליד הגיע מגול, אחר החברה, או פנה לשירות לסניף יתקבל הציון 1, אחרת 0.

לבסוף קיבלנו את דатаה פריטים הציונים הבא:

Scoring output - The rank (last) column

58...	df.groupby_segment_temp['rank']=df.groupby_segment_temp.sum(axis=1)
59...	df.groupby_segment_temp
60...	is_buisness_score department_score car_year_score platform_score age_score car_price_score desirable_rental_days_score time_catagor_score profit_score Market Cap_score rank
0	0 0 3.0 0 1.0 2.0 1.0 0 2.0 1.0 10.0
1	1 1 4.0 1 4.0 4.0 4.0 1 4.0 4.0 28.0
2	0 0 1.0 0 2.0 1.0 2.0 0 1.0 2.0 9.0
3	1 1 3.0 1 3.0 3.0 3.0 1 3.0 3.0 22.0

כאשר ניתן לראות שמדатаה פריטים סיכום התוצאות, קיבלנו דטה פריטים של ציוניים, כאשר האשכול השני בעמודת Rank (אשר מייצגת את הסכימה של כל התוצאות), קיבל את הציון הגבוה ביותר (28), האשכול הרביעי קיבל את הציון השני הגבוה ביותר (22), האשכול הראשון קיבל את הציון (10), והאשכול השלישי את הציון הנמוך ביותר (9).

לפי הציוניים, אפשר בזורה אינטואטיבית להבין, שהאשכול השני בגלל שהוא קיבל את הציון הגבוה ביותר יקבל את הקטלוג ליד "רותח", האשכול הרביעי יקבל את הקטלוג ליד "חט", האשכול הראשון יקבל את הקטלוג ליד "בינוי" והאשכול השלישי יקבל את הקטלוג ליד "קר".

11.1.5.1 הלוגיקה מאחורי ניקוד התוצאות

לפי סקירת הספרות שבעצמנו, הסקנו שליד עסק' שמעוני בסגירת חוזהليسינגן לחברה שהוא עובד (ולא לשימוש פרטי), שאם הליד הזה יסגר, הרוח הפוטנציאלי ממנו יהיה יותר גדול מאשר לאי פרטי שמעוני ברכב לשימוש פרטי.

לכן נתנו את הציון 1 אם הערך הcy שכיח באשכול הוא ליד עסק', ו-0 אם הערך הcy שכיח הוא ליד פרטי.

לגביה מחלקות שבהן עובד הליד, נתנו ציון יותר גדול לילדים שעובדים במחלקות של שיווק ומכירה, משאבי אנוש, ואדמיניסטרציה, כי בדרך כלל מחלקות אלה, הם המחלקות בחברות העסקיות שמתעסקות בסגירת חוזהليسינגן לחברה ולכן נתנו ציון 1 לאשכול שהערך הcy שכיח שלו הוא אחד מהמחלקות האלו, ו-0 לאשכול שהערך הcy שכיח הוא מחלקה אחרת.

לגביה שנת הרכב, ככל ששנת הרכב יותר גבוהה, ככל סכום המכירה הפוטנציאלי יותר גדול, וכך ככל שנת הרכב השכיחה באשכול יותר גדולה יתקבל ציון גבוהה, כדי להפיק מקסימום רוח לארגון.

לגביה פלטפורמת הנחיתה שמננה הגעה הליד, לפי סקירת הספרות ומחקר שבעצמנו, הבנו שלדים אשר מגעים מופיע נחיתה כמו מודעות מגוגל, או מאתר החברה אשר הביעו התעניינות ישירה ונכנסו לאתר והשאירו פרטיים, או לדיים שהתקשרו ישירות לسنיף, יש סיכוי מכירה גבוהה יותר מכיוון שהם הביעו התעניינות ישירה לנו התקבל הציון 1.

לעומת לדיים שבאו מפלטפורמת כמו אינסטגרם ופייסבוק, שהם בדרך כלל לדיים בעלי שיעור המרה נמוך יותר (לדיים פחות "רציניים") לנו התקבל הציון 0.

לגביה גיל הליד, ככל שגיל הליד נמוך יותר, יש סיכוי יותר גדול שהוא תקופה ארוכה יותר בחברה הנוכחיית, וככה שימור הלקחות עם החברה והlid יהיה יותר טוב. בנוסף, יוכל להמליץ לעוד אנשים על שירות הליסינג, لكن גם לגיל יש משקל בחלוקת הציון ואשכול שהgil המוצע נמוך יותר קיבל ציון גבוה יותר.

לגביה תקופת ההשכרה הרצiosa בימיים, ככל שהlid מעוניין בהשכרה לטוווח רחוק יותר, ככל הרוח הפוטנציאלי שלו יותר גדול, לנו האשכול קיבל ציון גבוה יותר, וכך אותו הדבר לגבי רווחיות החברה.

11.2 החלק השני – הלמידה המונחית

כפי שציינו בהקדמה, לאחר שהלוקוח השתמש בקובץ הילדים המקורי שביצעו בחלק הראשון ואנשי המכירות ניסו לבצע מכירות, הלוקוח מעלה את קובץ הילדים בפעם השנייה כאשר כעת אנו יודעים האם הילד הפרק למכירה.

כעת נוכל להשתמש באלגוריתמים של למידה מונחית, להשוות בנייהם, ולבחרור את המודל הטוב ביותר.

הסביר באופן כללי על אלגוריתמים של למידה מונחית אפשר לקרוא בסקירת הספרות בסעיף 3.5.

הסביר על מדריך הדיקן השונים של למידה מונחית ניתן לקרוא בסקירת הספרות בסעיף 3.6.

لتיאור מלא של אופן חישוב המתמטי של האלגוריתמים בסעיף 9.

לאחר שביצעו את תהליך ETL, וה-Pre-Processing, חילקו את הדטה לדטה סט אימון ודטה סט מב奸, הדטה מוקן להתקבל קלט באלגוריתמים המונחים השונים, נתחילה בלאפין את התוצאות של כל אלגוריתם.

11.2.1 שימוש ברגסיה לוגיסטי

Logistic Regression

```
In [34]: logmodel = LogisticRegression(random_state=100)
logmodel.fit(X_train,y_train)
predictions_logmodel = logmodel.predict(X_test)

In [35]: print(metrics.classification_report(y_test,predictions_logmodel))
```

	precision	recall	f1-score	support
0	0.87	0.92	0.89	203
1	0.54	0.40	0.46	47
accuracy			0.82	250
macro avg	0.71	0.66	0.68	250
weighted avg	0.81	0.82	0.81	250

יצרנו משתנה `logmodel` אשר הוא מודל רגסיה לוגיסטי אשר מובא מהספרייה Scikit-Learn. לאחר מכן השתמשנו בפונקציה `fit()`, אשר גורמת למודל להתאים על דטה סט האימון כולם על האיקסים שלנו (כאשר כל איבר ב-`X` מייצג ליד), ולמצוא את הקשר ל-`y`, מערך אשר מייצג האם הילד נמכר או לא.

יצרנו משתנה בשם `predictions_logmodel`, שלוקח את המודל אחרי שהוא התאים על הדטה, ועכשו אחורי שהוא למד באמצעות הפונקציה הסיגמואידית מה מאפין ליד שהפרק למכירה, אפשר לבצע חיזוי באמצעות הפונקציה `predict()` על דטה `X` שהם האיקסים שלנו כולמר הילדים, ללא המידע האם הילד הפרק למכירה.

מכיוון ששמרנו את התוצאות האמיתיות האם ליד הפרק למכירה ב-`X`, נוכן להציג מודל שנקרו `classification_report`, שישווה את התוצאות של הרגסיה הלוגיסטי לעומת התוצאות שקרו בפועל.

11.2.1.1 ניתוח התוצאות של רגסיה לוגיסטי

קיבלה תוצאה דיוק Accuracy של 82%, תוצאה הנחשבת טוביה ממשועות הדבר היא שהמודול הצלח 82% מהפעמים לחזות נכון臆יה ליד הפרק למכירה.

אך, המטריקה Recall, קיבלת את הציון 0.4 ו-Precision שקיבלה את הציון 0.54.

מכיוון שהדטה מכיל יותר ילדים שלא נמכרו לעומת הילדים שנקנו נמכרו (עמודת המטריה לא מאוזנת), מודל הרגסיה למד בצורה טוביה יותר לחזות לילדים שלא הרכו למכירה, ולכן הוא התקשה בחזות לילדים שנקנו הרכו למכירה.

לכן Recall של 40% אומר שמתוך הילדים שבאמת נמכרו, הוא הצלח לחזות בהצלחה רק 40% מהם, ואת השאר הוא לא הצלח לחזות, כלומר הוא חזה אותם כ-0.

-Precision של 54% אומר שמתוך הלידים שנמכרו, הוא חזה גם לידים אחרים שימכרו, שלא באמת נמכרו.

לכן ניתן להגיד שמודל הרגרסיה הלוגיסטי הצלח לחזות בצורה טובה מאוד לידים שלא נמכרו, ולא הצלח לחזות בצורה טובה כל כך את הלידים שהפכו למכירה.

11.2.2 שימוש בעז החלטה

Decision Tree

Decision Tree before using Cross Validation

```
In [36]: dt = DecisionTreeClassifier(random_state=101)
dt.fit(X_train,y_train)
predictions_dt = dt.predict(X_test)
```

יצרנו משתנה שנקרא dt (קיצור של decision tree) ויישמו בו את המודל sklearn.DecisionTreeClassifier, מודל בעז החלטה מספוריה של הפונקציה fit(), ולחזות על

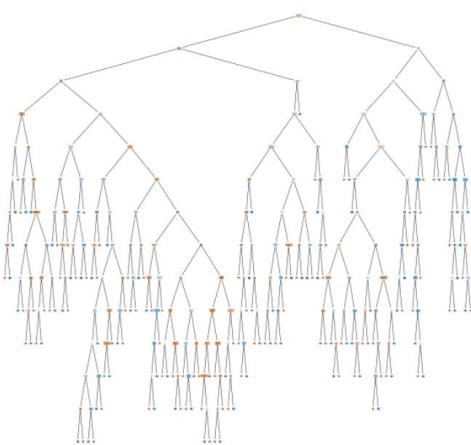
גם כן נתנו לעז ההחלטה להתאמן על דатаה סט האימון באמצעות הפונקציה predict(), ולחזות על דטהה סט המבחן באמצעות הפונקציה predict().predictions_dt

את התוצאות שמרנו במשתנה בשם predictions_dt

11.2.2.1 Grid Search בעז שימוש ב-CV

היתרון הגדול של עז ההחלטה, שהפלט שלהם הוא אינטואיטיבי למשתמש ונitin לראות בכל ענף, כיצד הפעילה ההחלטה ובסוף, לראות מה התנאים שהוא חזה שביהם ליד הפור למכירה. אילוסטרציה של העז ההחלטה שנוצר לנו לאחר שהוא התאמן על הדטהה סט נראה כ:

```
In [37]: classes = ['sold','not sold']
In [52]: plt.figure(figsize=(20,20))
features = df_lead_for_analysis.columns
tree = tree(dt, features, classes, class_names=classes, filled=True)
plt.show()
```



ניתן לראות שמכיוון שיש לנו הרבה פיצרים (עמודות) בדטהה סט, העז התפצל להרבה ענפים כאשר בכל שלב הוא חישב לכל עמודה את החומרה Information Gain וcut את העז נמצא במצב של Overfitting. למעשה, הוא התאים את עצמו יותר מדי לדטהה סט האימון, ובשביל לקבל תוצאה האם הליד הפור למכירה או לא, ישנו הרבה תנאים בשבייל שיתקבל סיווג.

לכן, בהמשך במודל של Cross Validation נגדיר את ההיפר פרמטרים ונמנע מהעז לגודל יותר מדי ובכך להפוך אותו לפחות יותר קרייא יותר ובעל אחוז דיוק גבוה יותר.

לכן, לשיכום, גרפ העץ לא קרייא בغالל שהוא התפצל ליותר מדי ענפים מה שיכול להעיד על תוצאות החיזוי שלו.

11.2.2.2 ניתוח התוצאות של עץ החלטה

```
: print(metrics.classification_report(y_test,predictions_dt))

precision    recall   f1-score   support
          0       0.85      0.74      0.79      203
          1       0.28      0.43      0.34       47
accuracy                           0.68      250
macro avg       0.56      0.58      0.56      250
weighted avg    0.74      0.68      0.71      250
```

ניתן לראות שכי שהנחנו שראינו כיצד נראה העץ, תוצאות המטריקות הן לא טובות, אחוז הדיקון הכללי Accuracy הוא 68%, אחוז דיקון שהוא לא מספיק גבוה בשבייל שנוכל להסתמך על המודל.

נשים לב שה-Precision בתוצאת חיזוי לדיים שהפכו למכירה היא 28% זהה נמוך מאד, ככלומר מתוך הלידים שנמכרו הוא חזה גם המון לדיים שלא באמת הפכו למכירה.

גם ה-Recall של ביחסו הלידים שנמכרו קיבל ציון נמוך של 43%. לשיכום, מודל עץ ההחלטה הפשטן ללא הגדרת ההיפר פרמטרים עשה עבודה לא טובה בתור מודל חיזוי.

11.2.3 שימוש בעץ החלטה יחד עם פונקציונליות של CV

Using Grid search on Decision Tree

```
In [53]: params = {'max_depth': list(range(1,10)),
               'min_samples_split': [2,3,4,5],
               'min_samples_leaf': [1,2,3,4,5]}

cv_dt = GridSearchCV(estimator=tree.DecisionTreeClassifier(criterion="entropy"),param_grid=params)
cv_dt.fit(X_train,y_train)
predictions_dt_cv = cv_dt.predict(X_test)
```

על המשמעות של Grid Search בلمידת מכונה הרחובנו בסעיף [9.6](#).

יצירת המשתנה params שהוא המשתנה שיגדיר את ההיפר פרמטרים של העץ, כאשר בכל פעם נבנה עץ אחר עם אחד מהפרמטרים ולבסוף יתקבל העץ עם אחוז הדיקון הגבוהה ביותר. ההיפטר פרמטרים שנבחרו לבניית העץ הטוב ביותר:

הסבר	היפר פרמטר
עומק המקסימלי שהעץ יוכל להגיע אליו, ככל שהעומק יותר גבול, כהה יהיה יותר תנאים שהleaf יצטרך לקיים בשבייל שהעץ יחולט אם לחזות שהוא הפר או לא הפר למכירה.	Max_depth
מספר הערכים המינימלי שצרכי להכיל feature מסוים בשבייל שהוא יוכל להיות צומת ההחלטה.	Min_samples_split
מספר הערכים המינימלי לאחר הפיצול שעלה צורך לקיים בשבייל שהוא יוכל להיות צומת ההחלטה.	Min_samples_leaf

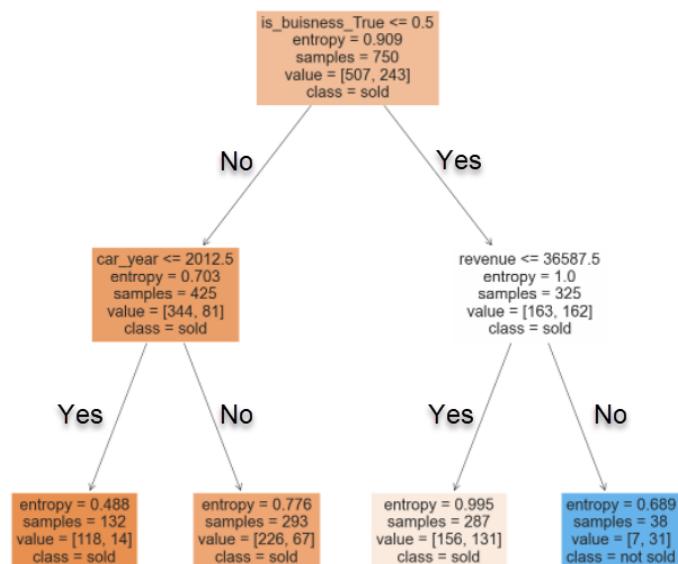
טבלה 11.2.3. הסבר היפר פרמטרים של עץ ההחלטה

נדגים את `Min_samples_leaf`, `Min_samples_split` ו-`Min_samples_leaf` לצורך המחשה.
 אם לדוגמה נרצה לקיים צומת החלטה האם הליד הוא מעל גיל 30 או מתחת.
 אם `the_min_samples_split` הוא 5, ויש 7 ילדים שהם מעל / מתחת לגיל 30, אז הצומת הזה יוכל להתקבל בעז.

אבל אם הגדרנו שה-`Min_sample_leaf` הוא 3, ולאחר צומת ההחלטה מתפצל בצורה שיש 6 ילדים שהם מעל גיל 30, ורק אחד שהוא מתחת לגיל 30, אז זה מבטל את הפיזול הראשון שהחליטנו כי אין לנו מספיק ילדים בשבייל לאפיון האם ה-`feature` של הגיל נותן לנו אינדיקציה אם ליד יכול להפוך למכירה. לכן, נסיק שהגדרת ההיפר פרמטרים אלה, ימינו מהען לצמוח להרבה צמתי החלטה כמו שקיבלנו בסעיף [11.2.2.1](#) וככה יפחיתו את הסיבוכיות שלו, ויגדלו את אחוז הדיווק. לאחר מכן הגדרנו משתנה שנקרא `cv_dt` (קיור של `Decision Tree`) (Cross Validation Decision Tree), ויישמו בו את המודל `GridSearchCV` שמקבל את סוג המודל המונחה אשר בחרנו, עץ החלטה, ואת היפר פרמטרים שהוא ירצו עלי, אשר בחרנו במילון `params` שהגדכנו. לאחר מכן נשתמש בפונקציה `fit()` אשר מתאמת על דатаה סט האימון. ומשתמשים בפונקציה `predict()` בשבייל לחזות על דטהה סט המבחן.

11.2.3.1 כיצד נראה העץ לאחר שימוש CV

```
plot_rf=cv_dt.best_estimator_
plt.figure(figsize=(20,20))
features = df_leads_for_analysis.columns
tree.plot_tree(plot_rf,feature_names=features,class_names=classes,filled=True)
plt.show()
```



נשים לב להבדל המשמעותי של כיצד נראה העץ לפני שהגדכנו את היפר פרמטרים בסעיף [11.2.2.1](#) ואחרי.

עכשו העץ הוא ייזואלי, קלומר ניתן לקרוא מה הם הצמתי ההחלטה שהתקבלו, לראות מה מدد האנטרופיה של כל אחד ומה ההחלטה שהתקבלה.

כפי שציינו ב-[9.4.5](#) ככל שהאנטרופיה יותר גבוהה כהה ממד האי וודאות נתונים שלוי יותר גבוהה (נתנו את הדוגמה של הטלת מטבע יש הסתברות של 0.5 לכל תוצאה, אך ממד האי וודאות שלוי הוגה והאנטרופיה תהיה מקסימלית ככלمر 1), ולהיפך, ממד ה-GI הוא ההפכי מאנטרופיה, קלומר כמה מידע אנחנו נקבל על ידי חלוקה לפי ה-`Attribute` זהה.

נשים לב שה-Root Note כולל צומת ההחלטה הראשי שהתקבל הוא האם הלוקו או עסק או לא, והאנטropיה שהתקבלה היא 0.9, מכיוון שרוח המידע שנקלט לפני הפיצול ואחרי הפיצול הוא הכי גבוה (כי רוח המידע מחושב על ידי חישור האנטרופיה לפני הפיצול שאצלנו הוא 0.9, ואנטרופיה המוצעת שנקלט לאחר הפיצול), לכן האם הלוקו עסק, קיבל משקל גדול בהחלטה והתקבל צומת הראשי של העץ.

אם הלוקו לא עסק, הוא עובר לצד שמאל, אם הלוקו כן עסק הוא עובר לצד ימין.

425 ערכים עברו לצד שמאל כאשר הם לקוחות לא עסקים לעומת 344 לקוחות עסקים שעברו לצד ימין.

ונתקד כרגע מצד שמאל, עכשו צומת ההחלטה האם 2012.5 גדול משלמת הרכב של הלידים, אם התשובה היא כן, הם יעברו לצד שמאל של העץ, ואם התשובה היא לא, הם יעברו לצד ימין (כלומר עליה השני משמאלי).

132 ערכים עברו לצד שמאל, ככלומר שנת הרכב שלהם קטנה מ-2012.5, ומ-293 עברו לצד ימין. מתוך 132 ערכים שהם גם לקוחות לא עסקים וגם שנת הרכב שלהם קטנה מ-2012.5, כ-118 מהם לא נמכרו, ו-14 מהם נמכרו.

לעומת העלה השני משמאלי, שהם לקוחות עסקים ושנת הרכב שלהם גדולה מ-2012.5, כ-226 מהם לא נמכרו, לעומת 67 שנמכרו.

כבר נוכל לשים לב בהפרש הגדול בין 2 העליים לעומת האם הליד הפרק למכירה (67 לדיים הפקו למכירה בעלה השני משמאלי לעומת 14 בעלה השמאלי), אנחנו יודעים לפי הדאטה סט שלנו שאם הליד הוא עסק, והוא מעוניין ברכבים בשנה יותר גבוהה, יש לצד סיכוי יותר טוב להפוך למכירה, העץ זיהה את "הדף" זהה, ועל פי כן הוא יוצר את מודל ההחלטה.

cut נתקד מצד ימין של העץ, כ-325 ערכים שהם לקוחות עסקים עברו לצד ימין, עכשו צומת ההחלטה האם הרוחות של החברות בהם הלידים עובדים קטן מ-36,587.5. אם התשובה היא כן, הם יעברו לצד שמאל של העץ (עליה השני ימין) ואם התשובה היא לא הם יעברו לעלה הכி ימני.

כ-287 ערכים עברו לעלה השני ימיין, ככלומר 287 ערכים הם לקוחות עסקים, שהrorות של החברות בהם הלידים עובדים קטן מ-36,587.5, מתוך כ-156 לדיים לא הפקו למכירה, לעומת 131 שנ-הפקו למכירה.

לעומת זאת, 38 לדיים עברו לעלה הכי ימני, ככלומר, 38 לדיים הם לקוחות עסקים שהrorות של החברות בהם הם עובדים גבוהה מ-36,587.5, מתוך כ-31 הפקו למכירה, לעומת 7 שלא הפקו למכירה.

גם כאן המידע הוא אינטואיטיבי, אשר אנחנו יודעים שאם הלוקו עסק, והחברה בה הוא עובד היא יותר רוחנית יש לו סיכוי גבוהה יותר להפוך למכירה, لكن עז ההחלטה זיהה את הדף זהה, וקבע שאלה יהיו צמחי ההחלטה החשובים של העץ.

11.2.3.2 ניתוח התוצאות של עז ההחלטה לאחר שימוש ב-CV Grid Search

In [62]:	print(metrics.classification_report(y_test,predictions_dt_cv))			
	precision	recall	f1-score	support
0	0.84	0.97	0.90	203
1	0.60	0.19	0.29	47
accuracy			0.82	250
macro avg	0.72	0.58	0.59	250
weighted avg	0.79	0.82	0.79	250

ניתן לראות שתוצאות הדיוק של עז ההחלטה הרבה יותר טובות לאחר שימוש ב-CV Grid Search מאשר לאחר שהגדכנו את ההיפר פרמטרים והגבילנו את גודלית העץ ואת קבלת צמחי ההחלטה. אחוז הדיוק Accuracy שקיבלנו הוא 82%, ככלומר הוא הצלח לחצות נכון 82% מדאטה סט המבחן.

אך נשים לב למטריקה Recall של חיזוי הלידים שנמכרו, שקיבל 19%, לעומת מtower הלידים שבאמת נמכרו, הוא הצליח לחזות רק 19% מהם, ואת השאר הוא חזה כ-0.60%. גם ה-Precision של חיזוי הלידים שנמכרו לא קיבל ציון גבוה במיוחד אשר קיבל את הציון 60%, לעומת מtower הלידים שבאמת נמכרו, הוא חזה 60% מהם ימכרו, ולידים שלא באמת נמכרו, הוא חזה שהם ימכרו לעומת מtower הוא חזה 0 כ-1.

11.2.4 שימוש ב-Random Forest

Random Forest

```
[1]: rfc = RandomForestClassifier(n_estimators=1500,random_state=6)
rfc.fit(X_train, y_train)
rfc_pred = rfc.predict(X_test)
```

השימוש ב-Random Forest הוא פשוט מבחינת הגדרה שלו, אך מה שקרה מאחורי הקלעים לא נגלה בעת שימוש במודול ואין אופציה לראות מבחן ויזואלית על פי מה התקבלה ההחלטה.

דבר ראשון אנחנו אונחנו מוגדרים משתנה שנקרא `rfc` (Random Forest Classifier) ואנחנו בוחרים את הפרמטר `n_estimators` להיות שווה ל-1500.

כלומר, יצירה של 1500 עצי החלטה, על פי כל אחד מהם, יוצר עץ עצמאי שלא תלוי אחד בשני, יחשיבו מה הפיצ'רים החשובים על פיהם תתקבל ההחלטה, ועל פי השכיח של ההחלטה של כל העצים, תתקבל התוצאה האם הליד ימכר למכירה.

11.2.4.1 ניתוח התוצאות של Random Forest

	precision	recall	f1-score	support
0	0.87	0.93	0.90	203
1	0.58	0.40	0.48	47
accuracy			0.83	250
macro avg	0.72	0.67	0.69	250
weighted avg	0.82	0.83	0.82	250

כפי שניתן לראות, אחוז הדיוק שקיבלו - Accuracy הוא הגבוה ביותר מבין כל המודלים שנבדקו. הוא הצליח לחזות נכון 83% מהילדים.

גם במודול זה, ה-Recall וה-Precision של חיזוי הלידים שיופיעו למכירה נמוך.

גם כאן הוא סופג את הבעיה שטמונה בדעתה לא מושן והמודול למד יותר טוב את הלידים שלא הפקו למכירה לעומת מtower האלה שכאן.

אך, במודול זה, לעומת כל המודלים, היחס בין ה-Recall ל-Precision הוא יותר גבוה מאשר השאר.

11.2.5 בחירת האלגוריתם הטוב ביותר ביותר בחלוקת זה נסביר איך נבחר את האלגוריתם הטוב ביותר ביותר בצורה אוטומטית.

Choosing the best model

```
[1]: def return_the_best_model(*args):
    df_metrics=pd.DataFrame(index=["logmodel","decision_tree","decition_tree_cv","random_forest"],columns=["Accuracy","Precision","Recall","F1-Score"])
    acc_list=[]
    precision_list=[]
    rc_list=[]
    f1score_list=[]
    for i in args:
        acc_list.append(metrics.accuracy_score(y_test,i))
        precision_list.append(metrics.precision_score(y_test,i))
        rc_list.append(metrics.recall_score(y_test,i))
        f1score_list.append(metrics.f1_score(y_test,i))
    df_metrics["Accuracy"] = acc_list
    df_metrics["Precision"] = precision_list
    df_metrics["Recall"] = rc_list
    df_metrics["F1-Score"] = f1score_list
    df_metrics["Sum"] = df_metrics["Sum"].sum(axis=1)
    df_metrics["Rank"] = df_metrics["Sum"].rank()
    print(df_metrics)
    return df_metrics[df_metrics["Sum"]==max(df_metrics["Sum"])]
```

יצרנו פונקציה שנקראת `return_the_best_model` אשר מקבל כקלט רשימה של אלגוריתמים, ומחזירה כפלט את האלגוריתם שスクל התוצאות שלו היו הטובות ביותר.

בתוך הפונקציה אנחנו יוצרים דאטה פריים חדש שנקרא `df_metrics` אשר האינדקסים שלו הם השמות של האלגוריתמים שביהם השתמשנו, והעמודות שלו הם כל המטריקות שעליהם התבוססנו **שהם: Accuracy, Precision, Recall, F1-Score**.

לאחר מכן אנחנו יוצרים רשימה של כל אלגוריתם, ובכל רשימה של אלגוריתם, אנחנו מחשבים בונפרד את כל המטריקות ומוסיפים אותה לרשימה.

לבסוף לפי סדר החישוב, אנחנו מוסיפים את המידע לדאטה פריים.

הדאטה פריים שייצרנו נראה כך:

```
df_metrics = return_the_best_model(predictions_logmodel,predictions_dt,predictions_dt_cv,rfc_pred)
```

	Accuracy	Precision	Recall	F1-Score	Sum	Rank
logmodel	0.824	0.542857	0.404255	0.463415	2.234527	3.0
decision_tree	0.684	0.277778	0.425532	0.336134	1.723444	1.0
decition_tree_cv	0.824	0.600000	0.191489	0.290323	1.905812	2.0
random_forest	0.832	0.575758	0.404255	0.475000	2.287013	4.0

ניתן לראות בצורה טבלאית את הסיכום של כל האלגוריתמים המונחים שביהם עשינו שימוש.

ישנם שיטות רבות לבחירת האלגוריתם הטוב ביותר, בעיקר במודל עסק, עדיף לארגן לבחור את המודל בעל ה-Precision הגבוה ביותר כי מה שחשוב הוא ללמידה בצורה גבוהה איזה לקוח הפרק מכך.

מכיוון שאנחנו רצינו להתחשב בכל הפרמטרים, כפרויקט תעשייתי, התחשבנו בכל המטריקות באופן שווה. אז יצירנו את עמודת ה-Sum, שהיא חיבור של כל המטריקות, והמודל שקיבל את הציון המבוקש ביותר בסה"כ, במקרה שלנו Random Forest, יתקבל כאלגוריתם הנבחר ויישמר בענין של הלקוח.

ນציג שהאלגוריתם הנבחר הוא Random Forest רק בדוגמה הספציפית הזאת, וגרסתה לוגיסטיבית היה קרוב מאוד בתוצאות הדיקוק שלו ל-Random Forest, ובאותה מידה עם דאטה אחר, האלגוריתם הזה או אלגוריתם אחר, גם יכול להיבחר.

לבסוף הפונקציה מחזירה את המודל עם הסכום בעמודת-h-Sum הכי גבוה, האלגוריתם שנבחר מועלה לענין של הלקוח, ובפעם הבאה שהוא משתמש במערכת, האלגוריתם ישר יוכל לחזות אילו לידים יפהכו למכירה.

11.3 החלק השלישי – חיזוי לידיים לפי המודל הטוב ביותר ביותר

כפי שפירטנו בסעיף הקודם, לאחר שבחולק השני מועלה לענין של הלקוח המודל למידה המונחית הטוב ביותר, כאשר הלקוח מעלה קובץ לידים חדש בפעם השלישי, נעשה שימוש במודל אשר חוצה אילו לידים יפהכו למכירה.

מכיוון שמדובר בלבד בלידים חדשים, אין לנו אינדיקציה אם החיזוי שלנו נעשה בצורה טובה או לא (כלומר אין לנו את עמודת המטריה), אבל אנחנו סיכום של אילו לידים לפי המודל הטוב ביותר (ש大概是 שלנו הוא הוא Random Forest) הפקו למכירה.

<pre> result_data.rename(columns = {'is_sold':'Total'}, inplace = True) df_groupby_segment=result_data.groupby('Result').agg({ 'Total':'count', 'is_business': lambda x: x.value_counts().index[0], 'gender': lambda x: x.value_counts().index[0], 'department': lambda x: x.value_counts().index[0], 'car_year': 'median', 'platform': lambda x: x.value_counts().index[0], 'age': np.mean, 'car_price': 'mean', 'desirable_rental_days': 'median', 'Market Cap':'mean', 'revenue':'mean', 'profit':'mean', 'Market Cap':'mean' }).reset_index() df_groupby_segment </pre>																																							
<table border="1"> <thead> <tr> <th>Result</th> <th>Total</th> <th>is_business</th> <th>gender</th> <th>department</th> <th>car_year</th> <th>platform</th> <th>age</th> <th>car_price</th> <th>desirable_rental_days</th> <th>Market Cap</th> <th>revenue</th> <th>profit</th> </tr> </thead> <tbody> <tr> <td>0</td> <td>716</td> <td>False</td> <td>Male</td> <td>Engineering</td> <td>2013.0</td> <td>Facebook</td> <td>56.923184</td> <td>11590.724120</td> <td>248.5</td> <td>10894.927793</td> <td>13771.097785</td> <td>325.449022</td> </tr> <tr> <td>1</td> <td>284</td> <td>True</td> <td>Female</td> <td>Marketing and sales</td> <td>2016.0</td> <td>Google</td> <td>48.151408</td> <td>35231.386514</td> <td>517.5</td> <td>106025.614085</td> <td>40753.657748</td> <td>4852.113028</td> </tr> </tbody> </table>	Result	Total	is_business	gender	department	car_year	platform	age	car_price	desirable_rental_days	Market Cap	revenue	profit	0	716	False	Male	Engineering	2013.0	Facebook	56.923184	11590.724120	248.5	10894.927793	13771.097785	325.449022	1	284	True	Female	Marketing and sales	2016.0	Google	48.151408	35231.386514	517.5	106025.614085	40753.657748	4852.113028
Result	Total	is_business	gender	department	car_year	platform	age	car_price	desirable_rental_days	Market Cap	revenue	profit																											
0	716	False	Male	Engineering	2013.0	Facebook	56.923184	11590.724120	248.5	10894.927793	13771.097785	325.449022																											
1	284	True	Female	Marketing and sales	2016.0	Google	48.151408	35231.386514	517.5	106025.614085	40753.657748	4852.113028																											

יצרנו דата פרויים חדש באמצעות אגרגציה כדי שיביצעו בחלק הראשון של הלמידה הבלטי מונחית בסעיף [11.1.4](#).

זה"כ המודל חזה ש167 מתוך 1000 לדים לא הפקו למכירה, רובם הם לקוחות לא עסקיים, המין השכיח בו הם גברים, שלרוב עובדים במלחמות של הנדסה, שנת הרכב השכיחה בה הם רכבים מ2013, פלטפורמת הנחיתה הפופולרית היא פייסבוק, הגיל הממוצע הוא 57, מחיר הרכב הממוצע שהילדים התעניינו בהם הוא 11,590 דולר, כמות הימים שבהם הם מעוניינים בהשכרה הם 248 ימים, שווי השוק של החברות בהם הם עובדים הוא 10K דולר, ההכנסות הממוצעות הם 13.7K דולר, והרווח הוא 325 דולר.

לעומת זאת, הוא חזה 284 לדים ימכרו, רוב הילדים שהוא חזה שיימכרו הם לקוחות עסקיים, המין הפופולרי הוא נשים, המחלוקת השכיחה שבבה עובדים ילדים הוא שיווק ומכירה, שנת הרכב השכיח הם רכבים מ2016, פלטפורמת הנחיתה הפופולרית הם לדים מגול, הגיל הממוצע של הילדים הוא 48, מחירי המכירות של הרכב בהם הם התעניינו הוא 35K, כמות הימים הרצiosa הממוצעת היא 517 ימים, שווי השוק הממוצע של החברות בהם הם עובדים הוא 106K, ההכנסות הם 40K, והרווח הממוצע של החברות הוא 4.8K.

לסיכום: ניתן לראות דמיון גדול בין החלק של הלמידה הבלטי מונחית, ובין החלק השלישי שבו השתמשנו באלגוריתמים של למידה מונחית, בסופו של דבר הילדים שהמודל חזה שיימכרו למוכרים הם לדים עסקיים ויוטר רוחניים, כדי שבחalk הראשון האלגוריתם הבלטי מונחיה קטלה את אותו סגנון של ילדים ליד "לוהט".

12 בדיקות והערכתה (System Testing and Evaluation)

12.1 תכנית בדיקות מערכת (STP)

תכנית בדיקות המערכת מצורפת לנוספחים בסעיף [17.2](#).

12.2 תרחישי בדיקות (STD)

תרחישי בדיקות המערכת מצורפים לנוספחים בסעיף [17.3](#).

12.3 דוח בדיקות מערכת (STR)

דו"ח בדיקות המערכת מצורף לנוספחים בסעיף [17.4](#).

12.4 הערכת המערכת המוצעת:

12.4.1 שלמות

המערכת שהקמנו شاملת מבחינת היעדים והמטרות שהצבנו לה. הצלחנו לבנות אפליקציית רשת על בסיס אתר, שהזזה לידיים ב-2 סוגים שונים של אלגוריתמים, תוצאות הדיק שקיבלונו היו טובות, ממשק האתר עובד בצורה טובעה.

על עמידה במדדים פירטנו תחת מטרות יעדים ומדדים בסעיף הבא [2.5](#), ועל שינויים שביצענו בפרויקט פירטנו בריכוז שינויים בסעיף [15](#).

1.4.2. אמינות (איכות)

למערכת שלנו יש 2 חלקים, שימוש באלגוריתם בלתי מונחה ובאלגוריתם מונחה.

באלגוריתם הבלתי מונחה הצלחנו לחלק את הלידים לפי אלגוריתם Clustering K-Prototypes Silhouette, ציון הלידים שקיבלונו הוא 0.27, אשר מגדד זה הוא טעון שיפור, אך הציון מושפע גם מכך שנרגלנו את הנתונים וגם מכך שהדעתה שלנו מכיל עמודות קטגוריאליות שבהם המטריקה לא התחשבה.

מבחינת האלגוריתם המונחה, הגיעו לאחוז דיק של 83%.

ראינו שהילדים שחזינו על קובץ הלידים החדש (שהאלגוריתם לא למד אותו), יש לננתונים קורלהציה ודמיון עם הלידים "הרותחים" שחזינו בשלב הקטלוג, וגם לילדים שבאמת הפכו למכירה בשלב השני.

יכלנו לשפר את האלגוריתם זה אף יותר והרחבנו על כך בפיתוחים עתידיים והמשך עבודתה [14.2](#).

1.4.3. עלות

היתרון הגדול של הפרויקט שהוא לא מצריך הרבה משבבים והfonקציונליות שלו פשוטה, لكن גם העלות יחסית זולה.

בשביל לשפר את האלגוריתם צריך עוד זמן פיתוח עבודה בהערכתה של \$2000.

עלות השימוש לענן היא 0.016 לג'גה בייט של זיכרון.

עלות ה-Domain הוא 20 דולר.

1.4.4. אבטחת מידע

אבטחת המידע שהממשנו דガש עליה היא בעת יצירת סיסמה, כאשר המשתמש יוצר סיסמה היא נשמרת כ-HASH לכומר כמספר מזהה ייחודי שאפשר להעתיק אותו, لكن גם יפרצו לדאטה ביבס, לא יוכל להתחבר לאתר המערכת.

הען שלנו גם מכיל Token ייחודי אשר רק באמצעותו אפשר לגשת ולבצע פעולות שונות.

1.4.5. ייחודיות ומקרויות

קיימות מערכות שונות לביצוע חיזוי של לידיים כמו מערכות של Salesforce, אך מה שמייחד את המערכת שלנו היא הפשטות שלה.

הליך לא צריך להטמע אותה במערכות הפנימיות של הארגון, מה שבדרך כלל לוקח הרבה זמן גם מבחינת הטמעה וגם מבחינת הזמן והסתגלות ללקוח.

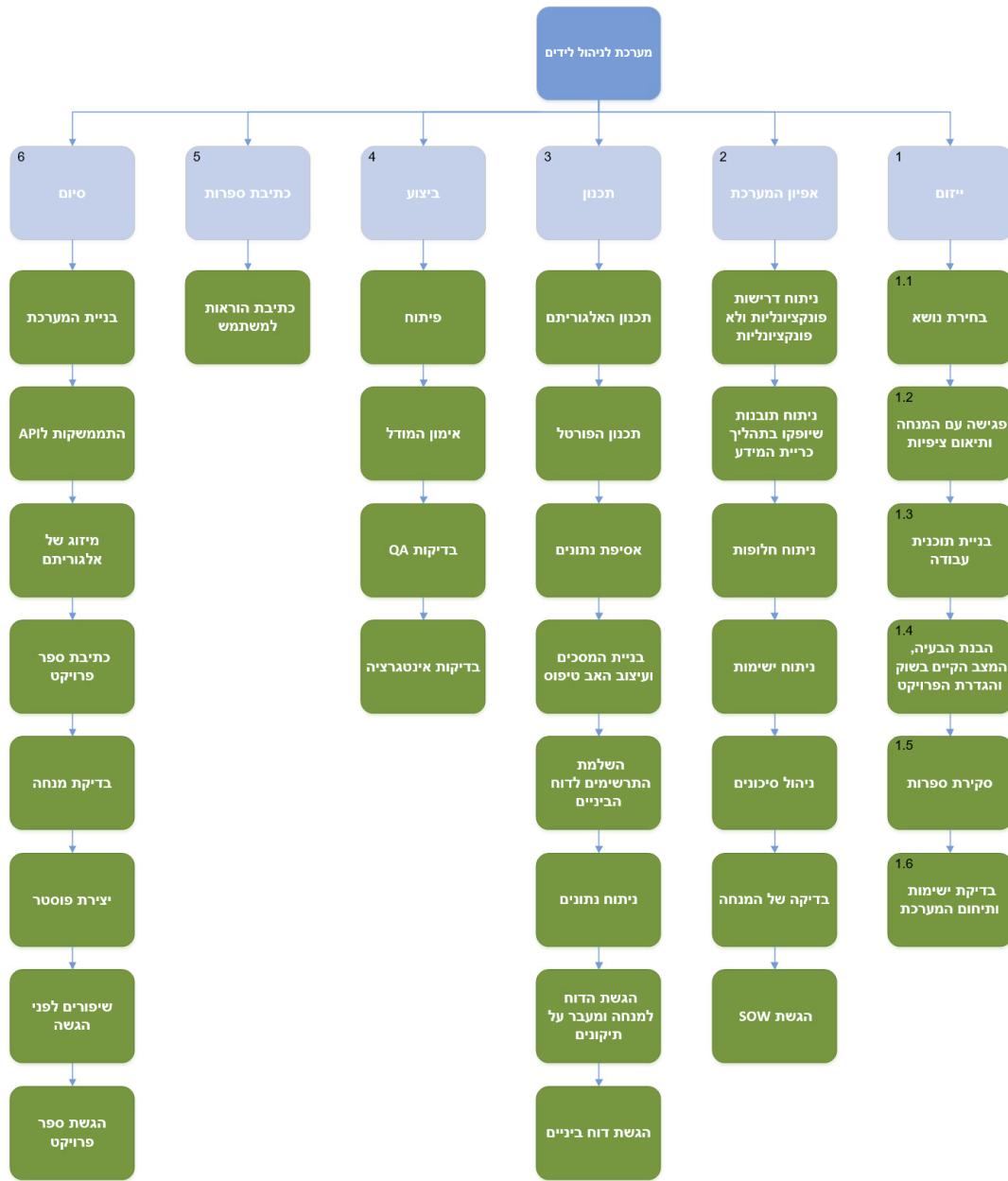
מדובר במערכת חיונית שモטעתה במיוחד לניהול לידים של חברות הליסינג, הליך צריך רק להעלות את הקובץ, ומבצעו סיווג באופן ישיר.

13 תכנית הפרויקט

13.1 תכנית עבודה. תרשימים גאנט מלא לכל מהלך הפרויקט

תרשים 13.1. תרשימים גאנט

תרשים 13.2 WBS



תרשים 13.2 WBS

תרשים WBS במסמך הבניינים לא כולל המשימות הכלולות בסיום הפרויקט - הוספנו את המשימות שנדרשו לפי התאריכים של תרשים הגאנט.

13.3 ניהול סיכונים

להלן טבלת הסיכונים המשמשים העולאים להיווצר בפרויקט. כל סיכון מקבל ציון בטוויח של 1-5 המבטא את חומרת הסיכון ואת הסבירות להתרחשותו.

שם הסיכון	הסיכון	תיאור	חווארה	סבירות (1-5)	פעילות לנטרול שבוצעו (1-5)	לוי"ז לביצוע
עירי ידע הדרושים להקמת המערכת	"יתכן כי יהיה קיימים ערים ידע הנובעים מחוסר ניסיון בפיתוח וכتنיבת אלגוריתמים	"יתכן כי יהיה עיר ידע הדרושים להקמת המערכת	5	4	התיעצות עם גורמים רלוונטיים ומהנהה המלאה להשלמת הפערים. למידה מעמיקה בקורסי התמחות או למידה עצמית של החומר הנדרש והתייעצות עם המנהה	החל מהagation SOW ועד סוף הפרויקט
התנגדות עובדים לשינוי המצב הקים	עובד חברות הליסינג בעליים להבע חשש מהטמעת התוסף בשל חשש מחידוש או קשיים טכנולוגיים.	עובד חברות הליסינג בעליים להבע חשש מהטמעת התוסף בשל חשש מחידוש או קשיים טכנולוגיים.	2	4	היות ולא עבדנו עם חברה ספציפית, לא התייחסנו לסיכון זה- אך בדקנו עם חברות ציבוריות אפשרויות הטמעה	בקרה לאורך הטעמיה התוסף
אי עמידה בלוחות זמינים, איחור בהagation מסמכים	עיכוב של אחד או יותר מאבני הדרך של הפרויקט.	עיכוב של אחד או יותר מאבני הדרכן של הפרויקט.	4	3	ניהול הפרויקט על פי תרשימים הגאנט המכיל בתוכו זמני תגובה לתרחישים לא צפויים, שיח קבוע בין חברי צוות הפרויקט	בקרה לאורך כל שלבי הפרויקט

	והצבת יעדים. עדכן שוטף מחלקת הפרוייקטים.- בוצע				
20.12.21	ליקחת קורס " ממשקי אדם מחשב", התיעצות עם איש מקצוע בתחום והמנחה- בוצע	1	3	ממשק לא ידידותי ולא אינטואיטיבי עבור המשתמש יכול לגרום לחשוף שיתופ פעולה מצד העובדים	בנייה ממשק לא ידידותי למשתמש
מתחלת הפייתוח וכתיבת הקוד	התיעצות עם המנחה וגורמים רלוונטיים, ניתוח האלגוריתם לעומק ושינוי בהתאם והתעמקות בנושא המודלים על כל סוג בסקירת הספרות- בוצע	2	5	פעולת אלגוריתם בצורה לא תקינה שלא تفسוג נכון את המידע הרצוי.	אלגוריתם לא תקין
מתחלת הפייתוח וכתיבת הקוד	קריאה של דוקומנטציה של המערכת בה נשתמש, ביצוע testים של העלאת קבצים לען לפני הקובץ האמתית התיעצות עם מומחים בנושא- בוצע	3	2	שימוש ב- API לטובות אחסון קבצים בענן באופן לא תקין	התממשקות לא נcona עם הענן

טבלה 13.3. ניהול סיכונים

14 סיום

14.1 סיכום ומסקנות

14.1.1 סיכום תהליכי הפרויקט

פרויקט מערכת ניהול לידים באופן חכם (Leadest) הינו פרויקט אשר כלל בתוכו שלבים רבים לטובות בניית המערכת.

תחילת ביצענו סקירת ספרות על עולם הבעה ועולם הלידים בפרט, למדנו מה מכיל ליד, כיצד חברות הליסינג עובדות ואילו פתרונות שונים ישנים בשוק.

לאחר מכן התחלנו לבash ולאפיין את המערכת אותה אנו רוצים לבנות, מה יהיו יכולותיה והפונקציונליות שלה. לבסוף, החלפנו שאנו רוצים לבנות מערכת פשוטה שתשתמש כ מוצר נוספת לסנייפ ליסינג ולא תדרש הטעמה במערכות החברה או שניי בשיטת העבודה שבה הם עובדים היום.

בשביל למש את הפרויקט, למדנו באופן עצמאי קורסים באתר הלימוד Udemy, קורסים של Flask, Data Science and Machine Learning וקורס פיתוח Web ב-.

למדנו על אלגוריתמים של למידה בלתי מונחת בשביל לבחור את האלגוריתם שמתאים ביותר לדאטה סט של הלידים לפני ניסיון מכירה, ולאחר מכן על אלגוריתמים של למידה מונחת, בשביל לבחור את האלגוריתם הטוב ביותר ביותר שיחזה איזה ליד יופיע למכירה.

בנוסף, הינו צריכים ללמד על תשתיות ענן של Google, חיבור באמצעות API למערכת Gmail לשילוח אימיילים ללקוח, ועל Database של SQLite בשביל למש את התשתיות של הפרויקט. לבסוף, הצלחנו למש את שלושת השלבים של הפרויקט, ותוצאות של 2 סוגים של האלגוריתמים עמדו במדדים שהגדרנו.

יצרנו דashboard ייעודי למנהל המכירות בשביל לנתח אילו לידים הפכו למכירה, וגם לנתח את הלידים באופן כללי לפני שאנשי המכירות יטסו לבצע מכירה.

2. בעיות שהתעוררו במהלך הפרויקט שהקשו על ביצועו

14.1.2.1 יצירת הדטה של קובץ לידים

בעיה מרכזית שעלה לנו במהלך הפרויקט, שמכיוון שלא עבדנו עם חברת ליסינג באופן ישיר, ולא היה לנו מקור נתונים של לידים, הינו צריכים ליצור קובץ לידים באופן עצמאי באמצעות Data Generator.

הבעיה הייתה שהוא השערכים שהוא צריך לכל ליד נוצרו באופן רנדומלי, ולא היה מאפיין ייחודי לקבוצה של לידים ולכן כל אלגוריתם בלתי מונחה שניסינו להפעיל על הדטה לא הצליח לחלק את הלידים לקבוצות בצורה טובה.

לאחר ביצוע מחקר מעמיק על סוגים שונים של לקוחות שמתעניינים בחזזה ליסינג, החלטנו להציג תנאים לייצרת הדטה עד שלבסוף, הדטה שנוצר לנו, מחקה בצורה טובה את התפלגות הלוקחות של חברת ליסינג, וגם מצאיך לקבל מידי דיק טוביים באלגוריתם של חילוק לאשכולות, וגם באלגוריתם המונחה.

14.2 פיתוחים עתידיים והמשך עבודה

בפרויקט זה פיתחנו מערכת על בסיס אפליקציית רשת, אשר יכולה לקבל קובץ לידים מה לקוחות ולבצע 2 סוגים שונים של סיוג, סיוג של למידה בלתי מונחת כאשר אנחנו מנוטים לקטלג מהו ליד בעל הסיכוי הגבוהה ביותר להופיע למכירה, וסיוג של למידה מונחת אחריו שהליך ביצוע ניסיון מכירה וcutout אנחנו יכולים ללמוד אילו לידים הפכו למכירה.

במהלך עבודה על הפרויקט ראיינו מקומות בהם אפשר לשפר את הפרויקט והפונקציונליות שלו.

14.2.1 עבודה עם חברת ליסינג וקבלת>Data

כפי שתיארנו בסעיף בעיות שהתעוררו במהלך הפרויקט [14.1.2.1](#) חלק מהבעיות שהקשו על מימוש הפרויקט היו שלא היה לנו DATA אמיתי מהלך כלומר מחברת ליסינג (כי לידים זה מידע רגיש שחברות ליסינג ששותחו עימם לא יכולו לשתף אותו), וקובץ הלידים שיצרנו התבסס על סקירת ספרות ומחקר שעשינו בראשת.

אם היינו עובדים עם חברת חיצונית והיינו מקבלים DATA אמיתי, היינו יכולים לבסס את האלגוריתמים שלנו אף יותר וללמוד אותם מtower DATA אמיתי.

14.2.2 שיפור האלגוריתמים ממפחחה המונחית

האלגוריתמים המונחיים שלמדו איזה DATA הפר למכירה הגיעו לאחוז דיוק (Accuracy) של 83% שזה נחשב אחוז דיוק טוב.

אך הם צלחו יותר בחזוי הלידים שלא נמכרו, מאשר הלידים שכן הפכו למכירה.

הבנו במהלך הפרויקט שהיא שפה עלי כרך שעמודות המטרה שלנו (האם הליד נמכר או לא) הייתה לא מדוונת, כלומר היא יותר לידיים שלא נמכרו מאשר אלה שנמכרו.

לפיכך עתידי היינו משתמשים בשיטות של Unbalanced Data בשבייל לשפר את ה-*Recall* וה-*Precision* של חיזוי הלידים שהפכו למכירה.

15 ריכוז שינוי מודוח התקן המפורט בפורמט טבלאי

מספר	השינוי	החלק בו בוצע	ביזמת	משמעות לפרויקט	מי השינוי
1.	שינוי מטרת הפרויקט	סטודנטים	חידדנו את מטרת הפרויקט ושינו את הנוסח	האב טיפוס הותאם למטרה החדשה, פירטנו על השינוי בסעיף תיקוף ובדיקות 6.2.2.1 .	
2.	האלגוריתמים	סטודנטים ומרצה	שינו את האלגוריתמים שהצרכנו שנעשה שימוש מודוח הביניים	האלגוריתם המונחה והבלתי מונחה הותאמו למידע שמכיל הלידים וגם לפט הרצוי.	
3.	התראות יעד שימוש במערכת	סטודנטים	החלנו לא למש את ביצוע מערכת התראות בפרויקט מכיוון שהבנו במלר דוח ביןיהם והדוח הסופי שהמערכת שלנו תהיה מערכת לשיווג ולא מערכת מודיע אשר תכיל פונקציונליות של התראה על ליד שלא מומש, لكن לא מימשו את יעד זה.		
4.	שינוי שחקן איש אדמיניסטרציה יה	סטודנטים	החלפנו את שחקן זה במנהל המכירות.	ב-SOW אמרנו ש מבחינת סדר הפעולות של התחברות למערכת, איש האדמיניסטרציה הוא זהה שיוסיף את הלידים ויעלה אותם לאתר, החלנו בדוח הביניים שייהו יותר נכון מבחינה היררכית שאיש אדמיניסטרציה יאוסף את הלידים ישלח אותם למנהל המכירות, והוא זהה שיעלה אותם למערכת, עדכנו זאת בתרשים ה- 7.2.3 Use Case Diagram	

טבלה 15. ריכוז שינויים

16 רשימת מקורות

1. (2022). Retrieved from Google Docs for Developers:
<https://developers.google.com/docs/api>
2. Algorithm, R. o.-V.-M. (2019). Research on K-Value Selection Method of K-Means Clustering Algorithm. *Graduate institute, Space Engineering University, Beijing*.
3. *Dash Enterprise*. (2022). Retrieved from <https://plotly.com/dash/>:
<https://plotly.com/dash/>
4. *Einstein Lead Scoring*. (n.d.). Retrieved from Salesforce:
https://help.salesforce.com/s/articleView?id=sf.einstein_sales_lead_insights.htm&type=5
5. *ELBAR Income statement 2021*. (2021). Retrieved from Maya Tase:
<https://maya.tase.co.il/reports/details/1438963/2/0>
6. Fabian Pedregosa, G. V. (2019). Scikit-learn: Machine Learning in Python.
7. HUANG, Z. (1997). CLUSTERING LARGE DATA SETS WITH MIXED NUMERIC AND CATEGORICAL VALUES. *CSIRO Mathematical and Information Sciences*.
8. Jeannette, P., & Matthew, W. (2020). Collaborative intelligence: How human and artificial intelligence create value along the B2B sales funnel. *Business Horizons*, 403-414.
9. *mockaroo*. (n.d.). Retrieved from Mockaroo Data Generator:
<https://www.mockaroo.com/>
10. Nusinovici, S., & Yih, C. (2020). Logistic regression was as good as machine learning for predicting major chronic diseases. *Journal of Clinical Epidemiology*, 56-69.
11. Nygård, R. (2019). AI-Assisted Lead Scoring. *Faculty of Social Sciences, Business and*, 78.
12. *pandas documentation*. (2021, December 12). Retrieved from Pandas:
<https://pandas.pydata.org/docs/>
13. Raza Hasan, S. P. (2018). Student Academic Performance Prediction by using Decision Tree Algorithm. *IEEE*.
14. Robert Nygård. (2020). Automating Lead Scoring with Machine Learning: An Experimental Study. *Proceedings of the 53rd Hawaii International Conference on System Sciences*, 1-10.
15. Ronacher, A. (n.d.). *Flask Documentation*. Retrieved from
<https://flask.palletsprojects.com/en/2.0.x/>
16. *scikit-learn Documentation*. (2021, December). Retrieved from <https://scikit-learn.org/>: <https://www.kite.com/python/docs/sklearn>
17. (2021). *Sixt Income statement*. SIXT. Retrieved from
https://irpages2.eqs.com/download/Companies/sixt/factsheet_271_english.pdf

17 נספחים

17.1 מאמר באנגלית

מאמר באנגלית נמצא בקובץ ZIP הפרויקט בתוך תיקיית ARTICLE.

17.2 תכנית בדיקות המערכת - STP

17.2.1 הקדמה

בנספח זה נפרט את תוכנית הבדיקות עבור מערכת ניהול לדיים, נבדוק את תקינות הירשומות באתר המערכת, התחרבות למערכת, עליית הקובץ לאתר, התממשקות עם שירות הענן, הצגת דאשborad, ומעבר בין הדפים השונים.

17.2.2דרישות הסביבה

17.2.2.1 מכשיר עם צג.

17.2.2.2 מכשיר עם חיבור לאינטרנט.

17.2.2.3 גישה לענן וחיבור עבור הקבצים שיעלו לאתר.

17.2.2.4 קובץ CSV המכיל לדיים (על מנת לבדוק את תקינות האלגוריתמים)

17.2.3 פירוט הבדיקות

17.2.3.1 בדיקת הממשק משתמש

מטרת הבדיקה:

לבדוק כי המערכת נוחה, פשוטה, ברורה ומובנת למשתמש, מבחינת חייה משתמש.

ויזוא עיצוב של האתר מבchinת פריסת המסך, כמו למשל שהרצלות מותאמת למסך ושל הדוחות עצםם, והתהליכיים במערכת רציפים ומובנים.

17.2.3.2 בדיקות התממשקות אינטגרציה

מטרת הבדיקה:

לאחר בדיקת כל רכיב על ידי בדיקת יחידה נבצע בדיקת תפקודם של כל המערכות ביחד נשותמש.

ביצוע שלילוב בין מרכיבי המערכת לבין הסביבה בה הם מיועדים לפעול, כמו למשל אויר המערכת עובדות עם התממשקות אחת לשניה, לדוגמה עליית הקובץ לענן ולראות שהוא מתקיים באתר המערכת.

מטרת בדיקה זו היא לא לוודא שהאלגוריתם עובד אלא שיש ממשק בין המערכות השונות, ושכל התהליך עובד באופן חלק.

17.2.3.3 ביצועים

מטרת הבדיקה:

בבדיקה זו נבדוק את הזמן שלוקח לקובץ לעלות לאתר, את זמן החישוב של האלגוריתם הבלתי מונחה בחלק הרាសן ושליחת אימיל ללקוח, בחלק השני כאשר האלגוריתמים של למידה מונחית לומדים את הלידים ונשמר האלגוריתם הטוב ביותר ביותר.

ובחלק השלישי כמשמעותו קובץ לדיים חדש, שימוש באלגוריתם הטוב ביותר משיכבה מהען, שימוש בקובץ לדיים החדש ולצורך חיזוי, ושליחת החיזוי באימיל ללקוח.

17.2.3.4 בדיקת אבטחת מידע

מטרת הבדיקה:

בדיקות הקשורות לאבטחת מידע דואגות להבטיח את שלמות ובטחון הילדים אשר מכיל מידע רגיש על ליקחות, בדיקות אלה יעזרו לוודא כי הנתונים הרגישים על הילדים לא יגעו לגורם זר ורק המורשים לכך יוכלו לגשת אל הנתונים כלומר רק מי רשום למערכת.

STD – Software Test Description מסמך עיצוב בדיקות המערכת

ס"ד	בדיקה	תרחיש	נתוני קלט	תוצאה צפואה
.1		הרשם למערכת – דף משתמש	תפקיד (מנהל מכירות, מנהל סניף), אימיל, שם משתמש, שם מלא, תעודת זהות, טלפון, מין, תאריך לידיה, סיסמה.	קלט תקין - ההרשמה הבוצעה בהצלחה והנתונים גרשמים במסד הנתונים. קלט לא תקין – המשמש הzin מידע לא על כר בהרשמה ותקוף לו על כר ההערה בשדה המתאים (לדוגמא שם משתמש תפוף, אימיל שכבר רשום במערכת וכו').
.2	בדיקות המשחק משתמש	הרשם למערכת – סניף	שם החברה, כתובת, עיר	קלט תקין - ההרשמה הבוצעה בהצלחה והנתונים גרשמים במסד הנתונים. קלט לא תקין – המשמש לא הzin קלט של שם חברה, תוחזר לו על כר הودעת שגיאה.
.3		התחברות למערכת	שם משתמש וסיסמה	קלט תקין – המשמש מעבר לתפריט הראשי. קלט לא תקין – העלאת שגיאת התחברות (שם משתמש לא תקין / סיסמה, משתמש לא קיים).
.4		העלאת הקובץ על ידי המשתמש	קובץ CSV המכיל רשומות לידיים.	קלט תקין - הودעה שהקובץ העלה בהצלחה, אימיל נשלח ללקוח. קלט לא תקין – הקובץ שגוי (פורמט קובץ לא נכון, קובץ גודול מדי, קובץ ריק).
.5		צפיה בדוחות	לחיצה באתר על צפיה בדוחות לאחר שהקובץ העלה ועובד.	הדווחות מוצגים למשתמש בצורה תקינה.
.6		העלאת הקובץ לען	קובץ CSV שמועלה דרך המערכת (לאחר לען)	קלט תקין – התחלה הפעלת האלגוריתם.

קלט לא תקין – החזרת הودעת שגיאה.	שהמשתמש העלה לאתר), "Token" ייחודי לאבטחת מידע.			
קלט תקין – הקובץ התקבל במלואו על ידי המשתמש. קלט לא תקין – הקובץ לא נשלח.	אימייל של המשתמש.	שליחת הקובץ המקטולג למשתמש.	7.	אינטרציה
זמן התגובה לא עולה על 3 שניות	Login	התחברות לאתר	.8	
קבלת הודעת חיבור הצלחת.	שם משתמש וסיסמה של אימייל החברה.	התmeshקות עם Gmail	.9	
לא עולה על 30 שניות (תלוî בגודל הקובץ וחיבור לאינטרנט).	קובץ CSV	העלאת הקובץ	.10	
1 דקוט		שליחת הקובץ המעודכן	.11	ביצועים
1 דקוט		ריצת האלגוריתם	.12	
5 שניות (תלוî חיבור האינטרנט של המCSIר).	הכנסת ה- <i>Domain</i>	העלאת האתר	.13	
20 שניות.	הקובץ הלידים המעודכן (לאחר שימוש בקובץ הראשוני).	הציג וчисוב הדוחות	.14	
לאחר 5 הזנות לא נכונות, המשתמש ייחסם ל-24 שעות.	שם משתמש וסיסמה לא נכונים או שאינם רשומים במערכת	עמידה בפני חדיות לא רצויות	.15	
כל חצי שנה, המשתמש יתבקש לשנות את הסיסמה לשיסמה בעלת 8 תוויים לפחות לפחות	שינוי סיסמה	ambio מידע	.16	abwechting

טבלה 10.3. בדיקות STP

17.4 דוח בדיקת אב טיפוס (STR)

שם הבדיקה: בדיקת תקינות האתר							
מספר	נושא הבדיקה	קלט	פעולה	תגובה רצiosa מהמערכת	עברית לא עבר	הערות	עברית
1.	כניסה למערכת		כניסה למערכת	העלאת האתר תוך פחות מ 3 שניות	עברית		
2.	הרשמה	פרטים אישיים ופרטית החברות הסניף	מעבר לעמוד הרשמה והזנת פרטים	העלאת העמוד תוך פחות מ 3 שניות והעלאת העמוד של הרשות הסניף	עברית	העמוד לוקח בצויה נכונה את מספר המזהה שלו אותו משתמש ומעבר לעמוד הרשות הסניף	
3.	הרשות סניף	פרטי הסניף	העלאת העמוד וטעינתו	העלאת תקינה בצויה	עברית		
4.	התחברות מייל וסיסמה	התחברות	העלאת עמוד התחברות הסניף להתחברות והציגת הודעה תודה על הרשות	מעבר מהרשות הסניף להתחברות והציגת הודעה תודה על הרשות	עברית		
5.	מעבר לעמוד הלקוח		העלאת העמוד	העלאת העמוד עם שם המשתמש שהתחבר	עברית		
6.	העלאת קובץ	בחירת קובץ CSV	אי הצגת שגיאה	הציגת הקובץ שהועלה	עברית		
7.	סיום האלגוריתם על הקובץ		הציגת עמוד תודה על העלאת הקובץ	והציגת עמוד מתאים בהתאם למספר הקובץ	עברית		
8.	כניסה לעמוד Dashboard		העלאת העמוד	העלאת העמוד עם שם המשתמש שהתחבר	עברית		

	עברית	הציגת קובץ שהועלה	אי הציגת שגיאה	בחירת קובץ CSV	העלאת קובץ ל- dashboard	.9.
	עברית	הציגת העמוד המתאים	הציגת עמוד לפי העמודה נמכר או לא	קובץ CSV	הציגת dashboard מתאים בהתאם לקובץ שהועלה	.10
	עברית	הציגת הגרף לפי בחירת המשתמש	בחירה בגרף	בחירת גרפ רוצי	הציגת הגרפים	.11

טבלת 17.4.

17.5 פוסטר

הפוסטר נמצא בתוך תיקיית POSTER בקובץ ההגשה של הפרויקט.

17.6 תיעוד ודפי נתוניים

17.6.1 קישורים למחברות פיתון שנכתבו במהלך הפרויקט

במהלך הפרויקט השתמשנו במחברות פיתון מסוג Jupyter Notebook על מנת לבצע את תהליכי ניתוח המידע (Data Analysis) לפני שתכתבו את הקוד להרצאה.

ניתן למצוא בקובץ ZIP הפרויקט תחת תיקיית NOTEBOOK_SCRIPTS את קבצי המחברת מסוג `dplyr`, וגם קבצי PDF על מנת שהיא תוכל לקרוא את הפליטים מבל' להריץ את הקוד עצמו.

17.6.2 תיעוד המידע בו עשינו שימוש

קבצי הלידים שבהם עשינו שימוש נמצאים בתוך קובץ ה-ZIP תחת תיקיית DATA.

17.7 מסמך ה-WoS המקורי

מסמך ה-WoS המקורי נמצא בתיקיית WOS בתוך קובץ ההגשה של הפרויקט.

17.8 נספחים נוספים

17.8.1 17.8.1.1 כל איסוף הנתונים ונתונים שנאספו

ראיון ראשון: פגישת התנהה יוסי ורוני – עובדים בחברת Hertz תל אביב.

איך העבודה נעשית היום?

רוני: היום העבודה נעשית בצורה ידנית, מנהל המכירות מקבל קובץ לידיים מאוחד הכלול את כל הפרטים אודוט היליד - בדרך כלל בימי ראשון ורביעי ושולח את הקובץ עם עמודה של אנשי המכירות שצרכים לפנות ליד. ואנחנו בתור אנשי המכירות פונים לאותו לקוח.

האם נעשה איזשהו תיעוד או דגשים מיוחדים לפני פניה ללקוח פוטנציאלי?

יוסי: אנחנו עוברים על פי הסדר של הליד שנכנס- הרាជון שיצר קשר הוא הרាជון שנטקשר אליו בלי תיעוד כלשהו. כמובן שיש עדיפות בין לקוחות שהוא פרטיל ללקוח עסקית אך אנחנו לא עושים איזשהו הבדלה.

האם נעשה ניסיון להטמעה מערכת CRM שמנהל בצורה חכמה לידיים?

רוני: לא שידוע לי, אבל שמעתי מחבר בחברה אחרת שהנושא היה מאייך ולא אופטימלי עבורם וכך הם נאלצו לוותר על כך בסוף.

כמה אנשי מכירות יש לכם בסניף והאם הם מוחלקים לאיזיהם תחומיים עיקריים שבהם הם עוסקים?

יוסי: לא, כולם בעלי אותו תחום וההבדל העיקרי בנינו הוא הותק. אנחנו 12 אנשי מכירות ועובדים כל יום במהלך השבוע.

האם יש יומם שבו אתם עוסקים יותר? מרגישים שלא יכולים לגשת ללקוחות?

יוסי: היום העמוס ביותר בשבוע הוא יום שישי, אנשים רבים מגיעים לסניף זהה ביום שאנו לעבוד מצאץ מצלחים להתקשר ללקוחות, ויש הרבה פעמים של לקוחות שייצרו אתכם קשר לאחר השארת הפרטים בשישי- הדבר כבר לא רלוונטי.

איך תרצו שהמערכת לניהול לידיים תראה?

רוני: חשוב מאוד שהמערכת תהיה יזואלית ונוחה למשתמש, אחרת לא יהיה שיתוף פעולה מצד המנהלים שלנו. בנוסף, חשוב שלא תעsha הבדלה משמעותית בחלוקת העבודה בין אנשי המכירות כי דבר זה יגרום לויכוחים בין חברי הצוות- בסופו של דבר, המכירות מהוות את מרבית השכר שלנו.

יוסי: אני חושב שצריך לעשות הפרדה בין לקוחות פרטיים ועסקים ושכן תהיה חלוקה בין אנשי המכירות שחלק יתעסק רק בפרטיים וחלק רק בעסקים- מדובר בזמן התעסוקות שונה וכן קרייטי עברו התנהלותם שלנו.

מה בנוגע להتمמשקות לענן בכך לשמר פרטיים קודמים?

רוני: אין שימור ידע יותר מדי בסניף, בדרך כלל לאחר יצירת קשר עם היליד- אנחנו מזינים את פרטי הלקוח למערכת והיליד נמחק. אולי אכן חשוב לשמר לידיים אחרים כי לא תמיד מצלחים לתפוס אותו ואנחנו מפספסים פה מכירה פוטנציאלית.

18 אב טיפוא

האב טיפוא עובד צורף לתיקיית הפרויקט ZIP תחת התיקייה SOURCE.