BUSINESS PROCESS INTELLIGENCE CHALLENGE

Data Science Project

By Moshe Habaz

תקציר

עם ההתפתחות המתמדת בעולם טכנולוגיית המידע, כמות המידע הדיגיטלי גדלה כל הזמן ועולה החשיבות של תחום כריית המידע. ניתוח בסיסי נתונים וחקירתו על ידי כלים אוטומטיים מאפשרים גילוי דפוסים אשר משמשים את מקבלי ההחלטות כדי לשפר תהליכים קיימים ובעלי משמעות עבורם.
בדוח זה אנו מציגים את ממצאינו מניתוח מאגר מידע המכיל את הליך הבקשות למתן סובסידיה עבור חקלאים גרמניים המשתייכים למשרד החקלאות האירופאי. מאגר מידע זה מכיל 2,514,266 אירועים עבור 43,809 בקשות על פני תקופה של שלוש שנים. המחקר נערך במסגרת BPI Challenge 2018.
יתר על כן, הניתוח מתבצע באמצעות שפת תכנות Python בשימוש אלגוריתמים ייעודיים לכריית נתונים כגון: Random Forest, Logistic Regression, Linear Regression ועוד.
ניתחנו היבטים שונים של התהליך בהתבסס על שאלות עסקיות בעלי עניין מרכזי לחברה וכתוצאה מניתוח מאגר המידע גילינו כי:

- קיים קשר מובהק בין אורך הבקשה, אי ציות לכללי הבקשה וגודל שטח החקלאי לבין דחייתה או אישורה. על ידי אלגוריתמי חיזוי מסוג קלסיפיקציה בנינו מודלים אשר חוזים זאת בהתאם למשתנים אלו ובדקנו את טיבם.
- במאגר קיימים קנסות שקיבלו מועמדים עקב חריגה מסוימת בהגשה, קנסות אלו נחלקים ל-2 קבוצות: "חמורים" או "קלים".
  כחלק מן ניתוח המאגר הערכנו באמצעות מודל פרדיקציה את גודל הקנס באמצעות סיווג מס' הקנסות "החמורים" ו"הקלים" עבור כל מועמד ועל ידי משתנים נוספים המאפיינים את החקלאי ואת החווה שלו.
- כל אירוע משויך ומטופל ע"י אחת מבין ארבעת המחלקות הבאות: e4,e7,d4,6b.
  עקב כך, זיהינו מי היא המחלקה העמוסה ביותר ומי הם המשאבים המנוצלים ביותר בה.

הקדמה

האיחוד האירופי מבזבז חלק גדול מהתקציב שלו על המדיניות החקלאית המשותפת (CAP). בין הוצאות אלה נכללים תשלומים ישירים אשר נועדו בעיקר לספק הכנסה בסיסית עבור החקלאים, הנפרדת מן הכנסה המושגת מהייצור. התהליך באתגר BPI 2018 מכסה את הטיפול בבקשות עבור תשלומים ישירים החל משלב הבקשה, ואם התהליך מאושר, עד לשלב התשלום של האיחוד האירופי עבור החקלאים הגרמניים החברים ב- European Agricultural Guarantee Fund. התהליך חוזר על עצמו מדי שנה בשינויים קלים עקב שינויים בתקנות האיחוד האירופי.
אופן העבודה הינו בעזרת מסמכים (סך הכל 9 מסמכים) המוצגים בטבלה הבאה כאשר כל מסמך כונס אל קובץ pickle :

שם המסמך	תיאור
Control summary	מסמך המכיל את סיכום התוצאות של בדיקות שונות.
Department control parcels (before 2017)	מסמך המכיל את אמיתות תוצאות הבדיקות של חלקות עבור מועמד יחיד.
Entitlement application	מסמך הבקשה לזכאות, כלומר, הזכות להגיש בקשה לתשלומים ישירים, שנוצרו בדרך כלל פעם אחת בתחילת תקופת מימון חדשה.
Inspection	מסמך המכיל את תוצאות הבדיקה באתר או בבדיקה מרחוק.
Parcel Document (before 2016)	המסמך מכיל את כל החלקות החקלאיות שעבורן ביקשו סובסידיה.
Geo Parcel Document	המסמך מכיל את כל החלקות החקלאיות שעבורן ביקשו סובסידיה. (המסמך החליף את המסמך Parcel document משנת 2016 ואת המסמך Department control parcels משנת 2017)
Payment application	מסמך הבקשה לתשלומים ישירים, לרוב מוגדר בכל שנה.
Reference alignment	מסמך המכיל תוצאות של סידור החלקות כפי שצוינו על ידי המועמד .

שאלות המחקר

1) האם בקשות שנדחו לוקחות יותר זמן מאשר בקשות שאושרו?
2) האם קיימת תלות לאורך שנים במתן קנסות עבור הבקשות ?
3) מהי המחלקה העמוסה ביותר ומיהו המשאב העמוס ביותר?

הפרויקט בנוי באופן הבא:
ראשית, נציג מידע על הנתונים השונים הקיימים במאגר אשר בהם השתמשנו בשאלות המחקר ,זאת בכדי להבין את סוג הקשר בין הנתונים, כגון: היסטוגרמה ומטריצת קורלציה ועוד.
שנית, נציג ניתוח מעמיק ומענה רחב על השאלות המרכזיות שצוינו לעיל.
לבסוף, נתאר את המסקנות העיקריות הנובעות מן שאלות המחקר.

תיאור מעמיק של הנתונים

כאמור, מאגר המידע מכיל 2,514,266 אירועים עבור 43,809 בקשות על פני תקופה של שלוש שנים. הבקשה הקצרה ביותר מכילה כ-24 אירועים, ואילו הארוכה ביותר מכילה 2,973 אירועים. בממוצע ישנם 57 אירועים לכל בקשה.
בקבצי ה- pickle של כל מסמך ניתן למצוא 76 עמודות הנחלקות ל-2: אלו המתארות את הבקשות ואלו המתארות את האירועים, כאשר כל רשומה מייצגת אירוע המשויך לבקשה מסוימת.
לצורך חקירת הdatabase נדרשנו להבין את שמות השדות : להלן 2 הטבלאות שמייצגות את המידע עבור הTrace ו- Event .
בטבלת ה-Trace ניתן למצוא משתנים המתייחסים לפרטי המועמד, מאפייני החווה שלו ופרטי הבקשה בשנה מסוימת.

בקשה-Trace		
שם השדה	**type**	**הסבר**
tr_department	literal	מס' מזהה של המחלקה
tr_application	literal	מס' מזהה של הבקשה
tr_year	literal	שנה שבה בוצע ה-trace
tr_number_parcels	discrete	מספר חלקות האדמה
tr_area	continuous	שטח כולל של החלקות
tr_redistribution	boolean	בקשה עבור הפצה מחדש של התשלום
tr_small farmer	boolean	בקשות של איזור קטן
tr_young farmer	boolean	בקשה לתשלום של חקלאי צעיר
tr_applicant	literal	מס' מזהה אנונימי למשתמש
tr_penalty_{xxx}	boolean	מסווג את סוג הקנסות בהתאם לסיבה מסוימת (JLP1, AVGP, C4, JLP3, JLP2, JLP5, JLP6, C9, V5, CC, AVUVP, GP1 ועוד)
tr_amount_applied{x}*	continuous	הסכום (ביורו) שהוגש בבקשה.
tr_payment_actual{x}*	continuous	התשלום שהתקבל על ידי המועמד (ביורו)
tr_penalty_amount{x}	continuous	סכום הקנס בהתאם לסיבות השונות
tr_cross_compliance	continuous	קנס על הפרה של ציות לכללים בביצוע פעולות
tr_rejected	boolean	האם הבקשה נדחתה?
tr_selected_risk	boolean	הבקשה נבחרה לבדיקה עקב הערכת סיכון

התכונות הבאות נרשמות עבור כל אירוע, כל האירועים נכללים בבקשות ועבור כל אירוע נשמר הזמן שבו הוא התרחש.

אירוע-Events		
שדה	**סוג**	**הסבר**
Eventid	literal	Id פנימי של האירוע
org:resource	literal	מציין את המשאב שאחראי לאירוע
time:timestamp	timestamp	הזמן שבו התרחש האירוע

*ישנם תכונות נוספות עבור הבקשות (כמו מס' מזהה עבור התוכנית אליה הבקשה שייכת, האם התשלום בסיסי , האם הבקשה נבדקה באופן אקראי או ידני ומזהים נוספים) ועבור האירועים (כמו שם האירוע , האם צלח , תת תהליך שבו הוא נרשם ועוד), אך מפני שלא נעשה בהם שימוש בניתוחינו אינם מצוינים בטבלאות אלו.

ביצוע תהליך ה- Pre-processing :

לאחר בדיקה וחקירת המסמכים עלה כי קיימת האפשרות שבקשה מסוימת תתפרס על פני יותר ממסמך אחד. לכן, מסיבה זו, על מנת לבדוק האם ישנו קשר בין אורכה של כל בקשה לבין דחייתה או אישורה ביצענו ריצה על כל המסמכים ואיחדנו אותם למסמך כולל מאוחד, אשר מכיל את כל רשימת האירועים ואורכו 2,514,266 שורות. לאחר סינונו של המסמך המאוחד, יצרנו Data frame מתאים עבור כל שאלה אשר קובץ על פי פרמטרים שונים כמו מועמד או בקשה , בהתאם לצורך.

בעת איחוד המסמך ביצענו שינויים בעמודות הבאות:

העמודות הבאות הומרו ממשתנים בוליאניים לערכי 1 עבור TRUE או 0 עבור FALSE:

- קנסות "קלים": ABP,AGP,AJLP,AUVP,AVBP,AVGP,AVJLP,AVUVP,B2,C4,C9,CC,GP1, JLP1,JLP2,JLP5,JLP6,JLP7.
- קנסות "חמורים": B3,B4,B5,B6,B16,BGK,C16,JLP3,V5,BGP,BGKV,B5F.
- tr_redistribution, tr_rejected, tr_selected_risk, tr_small farmer, tr_young farmer.

עמודת ה-time:timestamp הומרה לאובייקט מסוג datetime לצורך חישוב אורך הבקשה בעזרתה.

יצרנו עמודות נוספות:

- sum_light- עמודה זו מכילה את הכמות הכוללת של קנסות שסווגו כ"קלות" עבור כל רשומה הנמצאת במאגר.
- sum_worst- עמודה זו מכילה את הכמות הכוללת של קנסות שסווגו כ"חמורות" עבור כל רשומה הנמצאת במאגר.
- sum_amount_applied- עמודה הסוכמת את כלל העמודות של { }amount_applied.
- sum_penalty_amount- עמודה הסוכמת את כלל העמודות של { } penalty_amount.

בנוסף לכך, לצורך ביצוע הסכימה ערכים ריקים הומרו לערך 0.

נתונים סטטיסטיים:

בזמן חקירת שאלות המחקר נתקלנו במספר נתונים סטטיסטיים אשר תרמו בהבנה ולניתוח מעמיק של המאגר.

שם המסמך	כמות האירועים	אחוז האירועים שקיבלו קנס	כמות הבקשות	כמות המועמדים
Control summary	161,296	44.76%	43,808	15,937
Department control parcels	46,669	56.22%	29,297	15,388
Entitlement application	293,245	66.75%	15,260	15,086
Geo Parcel Document	569,209	26.17%	29,059	15,152
Inspection	197,717	66.97%	3,044	2,881
Parcel Document	132,963	75.11%	14,750	14,750
Payment application	984,613	37.09%	43,809	15,938
Reference alignment	128,554	51.14%	43,802	15,935

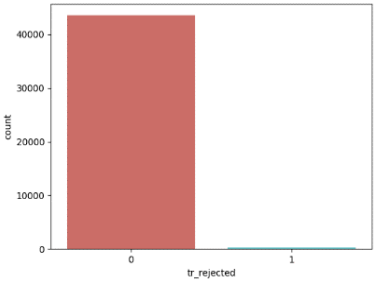
ניתן לראות כי במסמך ה-Payment application מופיעות כל הבקשות ושסך מגישי הבקשות הינו 15,938 מועמדים.

בנוסף, אחוז האירועים הגבוה ביותר שקיבלו קנס הינו במסמך ה- Parcel Document.

שם המחלקה	מס' האירועים בהם מטפלת	מס' משאבים במחלקה
e7	736,929	109
4e	735,363	58
6b	648,709	79
d4	393,265	60

מס' האירועים הגבוה ביותר מתקבל במחלקת e7 שבה גם מספר המשאבים הכי גבוה.

יתרה מזאת, כחלק מן הניתוח שבוצע רצינו לקבל התרשמות בנוגע למשתנים השונים כמו התפלגותם בתוך מאגר המידע. תחילה, נעשתה בדיקה על מספר הבקשות שאושרו מול הבקשות שנדחו על ידי במסגרת תכנית ה-CAP.
לאחר חלוקת המאגר לפי בקשות התגלה בפנינו כי ישנו חוסר איזון בין הבקשות שאושרו לבין אלו שנדחו.
ניתן לראות זאת בהיסטוגרמה הבאה שבוצעה על עמודת tr_rejected :



כמות הבקשות שאושרו:43,539 בקשות
כמות הבקשות שנדחו: 270 בקשות
חוסר איזון מובהק שכזה מהווה קושי במציאת תובנות נוספות לגבי דחייה או אישור הבקשה.

בנוסף, עבור כל ה- 43,809 בקשות קיבלנו את הנתונים הבאים (כאשר משתנה הduration מייצג את משך הטיפול בבקשה בימים) :

שם המשתנה	tr_area	tr_cross_compliance	duration
ממוצע	67.75	0.179621	334.808
סטיית תקן	83.30	1.747661	159.002
ערך מינימלי	0	0	113
ערך מקסימלי	576.07	100	1011

ניתן לראות שהתקבל פיזור רחב עבור משתנים אלו כלומר הנתונים אינם מקובצים סביב הממוצע אלא מפוזרים.

שם המשתנה	tr_rejected
הסתברות לדחיית הבקשה	0.006163
סטיית התקן	0.078264

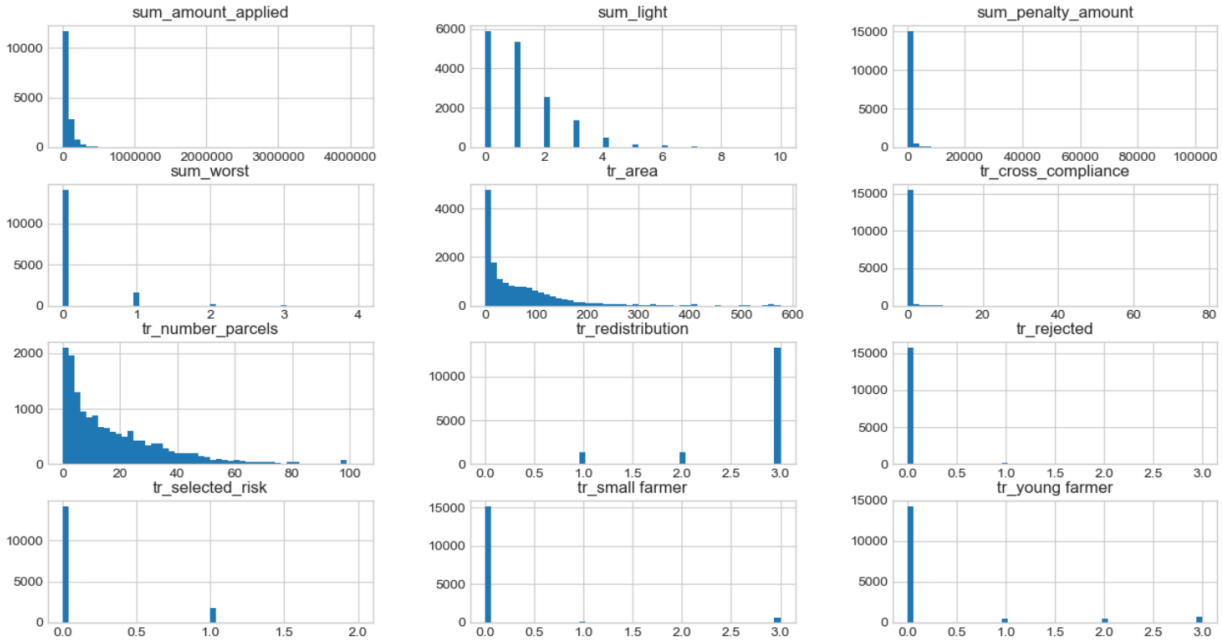
ניתן להסיק כי רוב הבקשות שנמצאות במאגר מתקבלות על ידי האיחוד האירופי.

לאחר חלוקת המאגר לפי מועמדים התקבלו התוצאות הבאות עבור פרמטרים המאפיינים אותו:

שם המשתנה	sum amount applied	sum worst	sum light	number parcels	cross compliance	area	young farmer	small farmer	selected risk	rejected	redistribution	sum penalty amount
ממוצע	78,305.34	0.1313	1.144	18.098	0.1786	66.512	0.223	0.1264	0.1134	0.0169	2.744	539.128
סטיית תקן	207,268.1	0.389	1.254	17.784	1.4186	82.458	0.7019	0.5833	0.3261	0.1539	0.605	2452.86
ערך מינימלי	0	0	0	0	0	0	0	0	0	0	0	0
ערך מקסימלי	4,110,014	4	10	103	78.333	576.071	3	3	2	3	3	102592.52

ניתן לראות כי סכום הקנסות המקסימלי שהתקבל עבור מועמד הינו 102,592.52 יורו וממוצע הקנסות עבור כלל המועמדים עומד על 539.128 כמו כן ניתן להסיק כי הפיזור רחב עבור משתנה זה דבר אשר יכול לעלות תהיות נוספות לגבי הערכת סכום הקנס הניתן למועמד.
כמו כן, מן הטבלה רואים כי הסכום המקסימלי מכלל הבקשות שחקלאי ביקש בשנים 2015-2017 סובסידיה עבור החווה שלו עומד על 4,110,014 יורו.

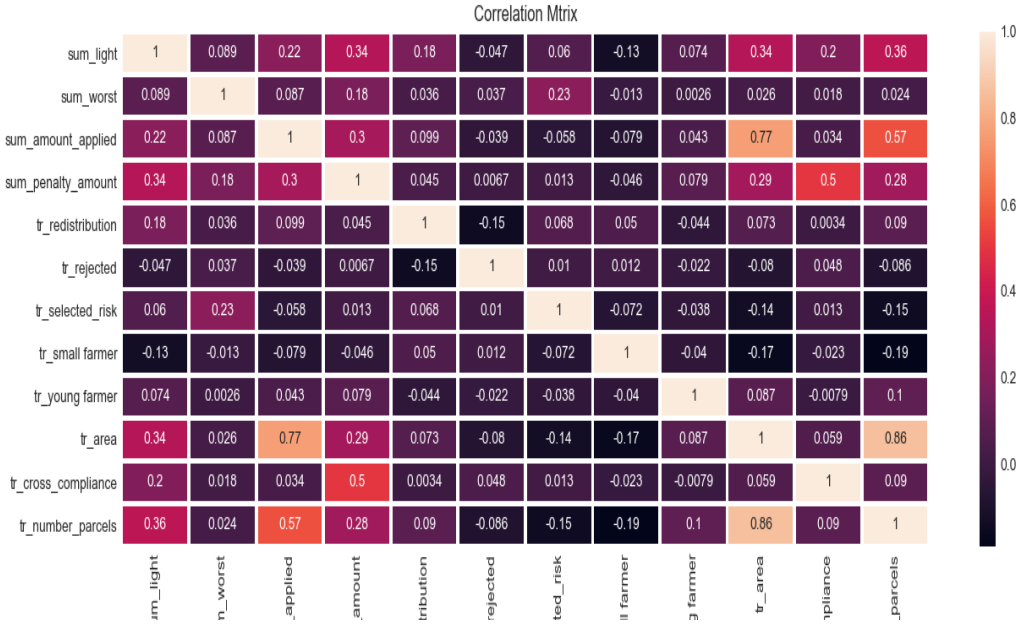
השטח הממוצע לחקלאי עומד על 66.512 הקטאר (יחידת מידה השווה ל-10,000 מטר רבוע).
בבדי לקבל ויזואליזציה לערכים אשר נותחו במאפיינים השונים של המועמד ביצענו היסטוגרמה לכל משתנה, תצוגה זו מאפשרת להראות את השכיחות היחסית שלהם ולקבל תפישה אינטואיטיבית ומהירה יותר של הנתונים.



מניתוח של חלק מן ההיסטוגרמות עלה:

- מרבית החקלאים אשר הגישו בקשה עבור סובסידיה הינם חקלאים מבוגרים ורק אחוז קטן מין הבקשות שייך לחקלאים אשר הגדירו עצמם כצעירים.
- אחוז רחב מן הבקשות אשר עתרו בקשה לסובסידיה הינם שייכים לחקלאים בעלי חוות גדולות ומעט מאוד מן הבקשות שייכות לבעלי חוות קטנות.
- מרבית הבקשות אשר נבחרו לבדיקה אינן נבחרו עקב הסיכון הקיים בהם.

ניתוח נוסף אשר ביצענו כחלק מן תהליך ניתוח מאגר המידע , היה בדיקת קורלציה בין המשתנים הנ"ל המאפיינים את המועמד, בבדי לדעת אלו משתנים קיים קשר ליניארי הדוק ובין אלו לא.



מן מטריצת הקורלציה ניתן להסיק כי בין גודל השטח לבין הסכום הכולל של הסובסידיה קיים מתאם חיובי בינוני חזק דבר אשר אינו מפתיע כיוון שאנו מצפים שבכל שגודל השטח גדול יותר כך הסכום שמופיע בבקשה יעלה.

ניתן לראות כי המתאם הגבוה ביותר מתקבל בין השטח לבין מס' חלקות האדמה ועומד על 0.88, מתאם זה הגיוני ומצביע על מולטי-קולינאריות בין שני מאפיינים אלו. בנוסף ישנו קשר בינוני בין מאפיין ה- cross compliance אשר מצביע על חריגה של המועמד בהגשת הבקשה לבין סכום הקנס הכולל שקיבל המועמד עבור הבקשה ועומד על 0.5.

שאלות המחקר:

כפי שצוין לעיל, בכלל השאלות איחדנו את כל 8 המסמכים למסמך מאוחד אשר עליו ביצענו את האנליזות השונות.

1. האם בקשות שנדחו לוקחות יותר זמן מאשר בקשות שאושרו?

לאחר שהתברר לנו כי בקשה מתפרסת על פני כל המסמכים בכדי למצוא את אורך הבקשה היינו חייבים לבצע הפרש בין הזמן שבו הבקשה מסתיימת לבין הזמן שבו היא התחילה וכך להגיע למשכה.

העמודות שנכללות בשאלה:

tr_applicant, tr_application, time:timestamp, tr_rejected, tr_area,tr_cross_compliance

בחרנו להתמקד בעמודות הנ"ל ממספר סיבות:

1. בעמודות tr_applicant, tr_application בחרנו כיוון שרצינו לבצע ניתוח נקודתי עבור כל בקשה של כל מועמד.
2. עמודת tr_cross_compliance אשר משקפת את אי ציות המועמד לכללי תכנית CAP, ערך גבוה במשתנה זה עלול לגרום לדחייה של בקשת הסובסידיה.
3. עמודת tr_area אשר נבחרה כיוון שיכולה להימצא אי התאמה בין הגודל המוצהר של חוות החקלאי לבין גודלה בפועל עשוי לגרום לדחיית הבקשה.
4. עמודת time:timestamp אשר נבחרה כיוון ששיערנו שבקשות שנדחו עשויות להימשך זמן ארוך יותר.

תיאור הניתוח:

- מתוך ה-D.B המאוחד ביצענו קיבוץ קבוצות לפי מועמד כמיון ראשי, ולפי בקשה כמיון משני.
- מכל אירוע המוכל בבקשה חילצנו את הערכים של tr_rejected, tr_area ו- tr_cross_compliance.
- כל בקשה נשלחה לפונקציה אשר בה הבקשות מוינו לפי פרמטר הזמן. על ידי מציאת ההפרש בין זמן האירוע שהתרחש אחרון לבין זמן האירוע שהתרחש ראשון גילינו את משכה של כל בקשה בימים.

לאחר מציאת משך הבקשה רצינו לקבל תחושה לגבי זמן הטיפול עבור כל בקשה, בין אם היא נדחתה או לא, ובכדי לבצע זאת ביצענו ממוצע לאורך הבקשות שאושרו ושנדחו:

ממוצע משך הבקשות שנדחו הינו : 358.3519 ימים.

ממוצע משך הבקשות שאושרו הינו: 334.6625 ימים.

ניתן לראות שבקשות אשר נדחו אורכות יותר זמן טיפול , מסקנה זו אינה מפתיעה כיוון שבקשות שנדחו עלולות לעבור בדיקות נוספות ושלבים אשר אינם חלק מהתהליך העסקי שעוברות בקשות שאושרו.

בעקבות הקושי אשר התגלה בשלב הניתוח הסטטיסטי בנוגע לכמות הבקשות הנדחות הקיימות במאגר המידע החלטנו לשנות את אופי השאלה ולבדוק האם קיים קשר בין אורכה של הבקשה, ערך הציות לכללי הבקשות וגודל שטח החווה לבין האפשרות כי הבקשה תידחה או תאושר.

בכדי לבצע ניתוח זה נעזרנו במודלים שונים של רגרסיה לוגיסטית, Random Forest ועץ החלטה המקבלים את 3 עמודות אלו כמשתנים מסבירים וחוזים לפיהם את המשתנה המוסבר tr_rejected .

במהלך בניית המודל נתקלנו בקושי נוסף כאשר במטריצת האמת ישנו חיזוי אבסולוטי של קבלת הבקשה דבר אשר לא נותן אינדיקציה כלשהי לחיזוי.

בכדי לתת מענה לאי איזון זה החלטנו לשמור על היחס הבא: על כל 20 בקשות שאושרו נרצה שיהיו 80 בקשות שנדחו. לכן, בחרנו אקראית ב- 2,160 בקשות שאושרו ושכפלנו את 270 הרשומות של הבקשות שנדחו כך שקיבלנו 540 דחיות ובסה"כ 2700 בקשות.

לאחר מכן, הקצאנו 60% מן הבקשות לתהליך של אימון (training) אשר משמש ללמידת המודל ולבחון את ביצועיו כלומר לבדיקת האלגוריתם, ו-40% מן הבקשות לתהליך של testing אשר מהווים מערך נתונים לבדיקת המודל מערך זה משמש להערכה בלתי מוטת של התאמה למודל הסופי על בסיס מערך האימון.

רגרסיה לוגיסטית-

ניתן לראות מפלט הרגרסיה הלוגיסטית כי כל המשתנים מובהקים מכיוון שכל המקדמים של המשתנים המסבירים (עמודת ה-coef) נמצאים בתחום של הרווח סמך (עמודות: [0.025 0.975]).
מקדמי הרגרסיה מצביעים על סבירות הניבוי של המשתנה המוסבר כאשר ערכים חיוביים מעידים כי התרחשות האירוע סבירה יותר וערכים שליליים מעידים כי סבירות האירוע נמוכה יותר . באופן כללי ככל שמקדם הרגרסיה קרוב יותר ל-0 הדבר מרמז כי ההשפעה שלו על המשתנה המנובא נמוכה יותר.

משוואת הרגרסיה הלוגיסטית שהתקבלה:

$$\log\left(\frac{p}{1-p}\right) = -0.9256 - 0.1118 \cdot area + 0.141 \cdot crossCompliance + 0.036 \cdot duration$$

```
                           Logit Regression Results
==============================================================================
Dep. Variable:            tr_rejected   No. Observations:                 1620
Model:                          Logit   Df Residuals:                     1616
Method:                           MLE   Df Model:                            3
Date:                Fri, 10 Aug 2018   Pseudo R-squ.:                  0.3565
Time:                        09:34:31   Log-Likelihood:                -521.63
converged:                       True   LL-Null:                       -810.65
                                        LLR p-value:                 5.778e-125
=====================================================================================
                        coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------
const                -0.9256      0.192     -4.813      0.000      -1.303      -0.549
tr_area              -0.1118      0.010    -11.227      0.000      -0.131      -0.092
tr_cross_compliance   0.1410      0.070      2.016      0.044       0.004       0.278
duration              0.0036      0.001      6.429      0.000       0.003       0.005
=====================================================================================
```

באשר p היא ההסתברות שבקשה תידחה.
Pseudo R-square- מדד המקביל למדד ה- R square הקיים ברגרסיה הליניארית, מציין כמה המודל מותאם לנתונים. הערך 0.3565 נמצא בטווח 0.2< 0.3565 < 0.4 ומציין כי קיימת התאמה.

```
>>>>>>>>>>>>>>>>>>>>>>>>>>LogisticRegression<<<<<<<<<<<<<<<<<<<<<<<<<<
             precision    recall  f1-score   support

          0       0.86      0.97      0.91       864
          1       0.78      0.36      0.49       216

avg / total       0.84      0.85      0.83      1080

                     Predicted Not Rejected  Predicted Rejected
Actual Not Rejected                     842                  22
Actual Rejected                         139                  77
mean accuracy on the given test data:  85.0926
```

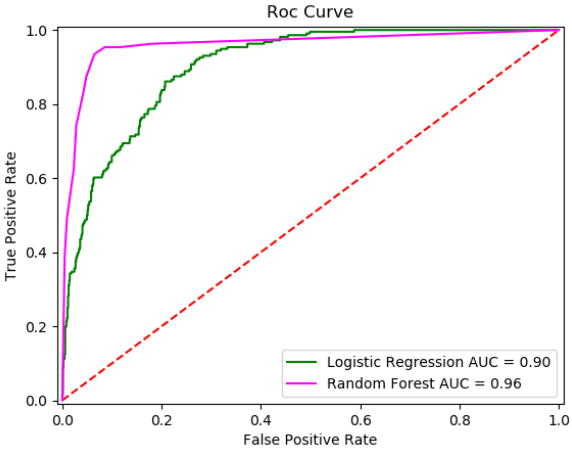
```
>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>RandomForest<<<<<<<<<<<<<<<<<<<<<<<<<<<<<
             precision    recall  f1-score   support

          0       0.96      0.95      0.96       864
          1       0.82      0.83      0.82       216

avg / total       0.93      0.93      0.93      1080

                     Predicted Not Rejected  Predicted Rejected
Actual Not Rejected                     825                  39
Actual Rejected                          37                 179
mean accuracy on the given test data:  92.963
```

בפלט שלעיל קיימים מדדים אשר מצביעים על טיב המודלים Random Forest וLogistic Regression,
נדגים את חישוב המדדים שהתקבלו באמצעות מודל Logistic Regression באופן הבא :
Precision דיוק הניחוש: אחוז הבקשות שניחשנו שיתקבלו ובפועל יתקבלו (t-pos=842) מתוך סך כל הניחושים שלי שיתקבלו (842+139).אחוז הבקשות שניחשנו שידחו ובפועל ידחו (t-neg=77) מתוך כל סך הניחושים שלי שידחו (22+77).
Recall: מדד הרגישות, אחוז הבקשות שניחשנו כי יתקבלו (t-pos=842) ובפועל באמת התקבלו (pos=842+22) ועומד על 0.97, אחוז הבקשות שניחשנו כי ידחו (t-neg=7777) ובפועל באמת נדחו (139+77) ועומד על 0.36.
f-score: המדד מחושב מתוך שקלול של דיוק הניחוש (Precision) והרגישות (recall).

השוואה בין המודלים:

ניתן לראות מן הפלט שהמודל Random Forest מניב תוצאות טובות יותר מן מודל הרגרסיה בכל המדדים שהוזכרו לעיל עבור ערכיו של משתנה החיזוי.
השוואה נוספת התבצעה באמצעות פונקציית ה- score אשר מעידה על הדיוק הממוצע המתקבל מתהליך ה- testing , לאחר ביצוע פונקציה זו על שני מודלי הניבוי התקבלו הערכים הבאים: עבור Random Forest התקבל דיוק של 92.963% ולעומת זאת במודל הרגרסיה התקבל ערך של 85.0926%.

בנוסף , לאחר שימוש באלגוריתמים השונים, בדקנו את עקומת ה - ROC על ערך הבקשות שנדחו.
עקומת ROC כוללת רק שניים מהמדדים במטריצת האמת: שיעור true-pos בציר האנכי, ושיעור false-pos בציר האופקי עבור חיזוי ערך הבקשות שנדחו. שיעור ה-true-pos מוגדר כאחוז המקרים בהם חזינו שהבקשה תידחה וצדקנו בתחזית, ושיעור ה-false-pos מוגדר כאחוז מהמקרים בהם חזינו כי שהבקשה תידחה וטעינו בתחזית.



כאשר רצינו לשלב את עקומת ה- roc של מודל עץ ההחלטה ראינו שצורתה אינה משקפת את הצורה האידיאלית של עקומת הroc. הגענו למסקנה שמודל העץ עושה overfitting למאגר המידע, כלומר, הוא מצליח למצוא קריטריונים שעושים הפרדה מלאה בין המשתנים והוא מותאם למדגם הנוכחי בלבד (הכוונה היא שיש לו וודאות מלאה בחיזוי). במקרה שכזה אין משמעות לעקומת ה-roc של מודל זה ולכן ויתרנו עליו.

מגרף העקומה המאוחדת של 2 המודלים ניתן לראות שהאלגוריתם האופטימאלי הוא Random Forest שבו צורת העקומה אידיאלית, כלומר הגבוהה ברוב התחומים ושטחה גבוה יותר (area-under-curve=0.96).
מעבר לכך, קיים חיתוך בין העקומות של המודלים ב false-positive-rate=0.45 ולכן עד ערך זה מודל ה Random Forest עדיף על הרגרסיה ונותן תוצאות טובות יותר אך מערך זה ומעלה נעדיף את אלגוריתם Logistic Regression שנותן ערך מעט גבוה יותר.

<u>מסקנות מן הניתוח:</u>
מהמחקר שביצענו עבור שאלה זו ניתן להסיק כי ישנו קשר מובהק בין אורך הבקשה, אי ציות לכללי שליחת הבקשה ושטח החווה של המועמד לבין דחייתה של הבקשה או אישורה.
בנוסף, את מודל ה- Random Forest אשר נבחר כטוב ביותר, ניתן להציע לאיחוד האירופי כמודל חיזוי עבור הסיכוי שבקשה ספציפית של מועמד תדחה או תאושר.
החיזוי מאפשר לצמצם את השלבים בתוך התהליכים העסקיים שהבקשה תעבור ולהפחית את הקצאת המשאבים עבור הבקשה בהתאם לכך ובמקרים מסוימים אף לחסוך בעלויות.

2. <u>חלק א- האם קיימת למידה בקרב בקשות המועמדים בשנת 2016 ובשנת 2017:</u>

<u>העמודות שנכללות בשאלה:</u>
sum_penalty_amount,tr_applicant,tr_application,tr_year
בחרנו להתמקד בעמודות הנ"ל ממספר סיבות:

1. בעמודות tr_applicant, tr_application בחרנו כיוון שרצינו לבצע ניתוח נקודתי עבור כל בקשה של כל מועמד.
2. עמודת sum_penalty_amount- מייצגת את הקנס הכולל אשר קיבל מועמד על חריגה כלשהי בבקשתו.
3. עמודת tr_year- אשר מייצגת את השנה הנוכחית בה הגיש המועמד את בקשתו.

תיאור הניתוח:

- ביצענו סינון לפי שנת 2016 ולפי שנת 2017 מתוך מאגר המידע המאוחד על מנת לקבל שני data frame שונים.
- כדי למנוע כפילויות כינסנו את המאגר לקבוצות על פי מועמד כמיון ראשי ואפליקציות כמיון משני , מכל קבוצה משכנו את השורה הראשונה (בעזרת פונקציית first).
- עבור כל מועמד סכמנו את הסכום הקנסות הכולל שהתקבל מכל הבקשות הקיימות לו בשנת 2016 עבור ה data frame של 2016 , תהליך זהה בוצע עבור שנת 2017.
- חיברנו את שני ה- data frame אל data frame אחד המכיל את הקנס הכולל עבור הבקשות של כל מועמד באותם השנים.
- חישבנו ממוצע לכל עמודה עבור שנת 2016 ושנת 2017.
- יצרנו עמודה חדשה שבה ביצענו הפרש בין סכומי הקנסות של כל מועמד בין 2016 ל-2017.

- על מנת לקבל תוצאות משקפות לא הכללנו מועמדים אשר לא קיבלו קנס באותם שנים.
- את עמודת ההפרשים הכנסנו לפונקציה המחשבת את טווח הרווח סמך בכדי לגלות האם קיימת למידה בקרב המועמדים.
- חישבנו קורלציה בין 2 העמודות המעידה על קשר בינוני בינהם.

```
average of panelty per applicant in 2016:  665.0801
average of panelty per applicant in 2017:  596.1823
Confidence interval: [ 84.5249 176.0057]
correlation:
  [[1.         0.35734311]
 [0.35734311 1.        ]]
```

<u>מסקנות מן הניתוח:</u>

ניתן לראות שממוצע הקנסות בשנת 2016 גבוה יותר מאשר הממוצע בשנת 2017 .
בנוסף, עבור עמודת ההפרשים התקבל רווח סמך אשר ב95% רמת ביטחון מכיל טווח של ערכים חיוביים בלבד דבר המצביע על השיפור שחל, כלומר, סכום הקנסות בקרב המועמדים קטן בשנת 2017.
תוצאות אלו מעידות על אפשרות של למידה בקרב המועמדים אשר קיבלו קנסות בעבר וכתוצאה מקנס זה משתפר אופי שליחת הבקשה לסובסידיה אצלם.

<u>חלק ב- חיזוי סכום הקנסות</u>

<u>העמודות שנכללות בשאלה:</u>

sum_light, sum_worst, sum_amount_applied, sum_penalty_amount, tr_applicant, tr_application,tr_redistribution, tr_rejected, tr_selected_risk, tr_small farmer, tr_young farmer, tr_cross_compliance ,tr_number_parcels,tr_area

בחרנו להתמקד בעמודות הנ"ל ממספר סיבות:

1. בעמודות tr_applicant, tr_application בחרנו כיוון שרצינו לבצע ניתוח נקודתי עבור כל בקשה של כל מועמד.
2. עמודת sum_penalty_amount- מייצגת את הקנס הכולל אשר קיבל מועמד על חריגה כלשהי בבקשתו .
3. שאר העמודות נבחרו כפרמטרים נוספים העשויים להשפיע על גודל הקנס.

תיאור הניתוח:

- שלפנו מתוך מאגר המידע את העמודות הרלוונטיות לשאלה זו.
- כדי למנוע כפילויות כינסנו את המאגר לקבוצות על פי מועמד כמיון ראשי ואפליקציות כמיון משני , מכל קבוצה משכנו את השורה הראשונה (בעזרת פונקציית first).
- יצרנו 3 data frame : עבור הקנסות, עבור המשתנים הבוליאניים ועבור שאר המשתנים הנומריים.
- ביצענו קיבוץ לפי מועמד ולאחר מכן:
  - סכמנו את הפרמטרים הבוליאניים השונים שמוכלים בבקשותיו.
  - סכמנו את עמודת sum_penalty_amount עבור כל הבקשות של כל מועמד.
  - ערכנו ממוצע לפרמטרים הנומריים השונים שמוכלים בבקשותיו.
- איחדנו את 3 ה-data frame אל data frame אחד בהתאם למשתנה tr_applicant המופיע בשלושתם.

<u>רגרסיה ליניארית</u>

בכדי לענות על שאלת המחקר בחרנו להשתמש במודל חיזוי מסוג רגרסיה ליניארית כאשר הגדרנו את עמודת sum_penalty_amount כמשתנה המוסבר ואת שאר העמודות כמשתנים המסבירים.
בכדי לבצע תהליך של אימון המודל (Training) הקצאנו 75 אחוז מן כלל הנתונים הקיימים ב-Data Frame המאוחד.
יתר נתוני ה-Data Frame עברו תהליך של testing לתיקוף המודל על בסיס תהליך האימון.
לפי מטריצת הקורלציה אשר מוצגת בסעיף נתונים סטטיסטיים ניתן לראות כי בין עמודות tr_num_of_parcel ו-tr_area קיימת מולטי-קוליניאריות העלולה לפגוע ביכולת החיזוי ובאיכות התוצאה של המודל הרגרסיה הליניארית.
בכדי לטפל בכך, לא כללנו את עמודת tr_area במודל.
לאחר הרצת המודל עם כלל המשתנים המוסברים קיבלנו שהמשתנים הבאים tr_redistribution, tr_small farmer, tr_rejected אינם מובהקים. עקב כך, החלטנו להסיר את המשתנים הללו ולהריץ את הרגרסיה בשנית על מנת לקבל רגרסיה הכוללת משתנים מובהקים בלבד.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     sum_penalty_amount   R-squared:                       0.387
Model:                            OLS   Adj. R-squared:                  0.386
Method:                 Least Squares   F-statistic:                     728.9
Date:                Fri, 10 Aug 2018   Prob (F-statistic):               0.00
Time:                        10:56:11   Log-Likelihood:                -74744.
No. Observations:                8106   AIC:                         1.495e+05
Df Residuals:                    8098   BIC:                         1.496e+05
Df Model:                           7
Covariance Type:            nonrobust
=====================================================================================
                        coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------
const             -1233.9176     59.662    -20.682      0.000   -1350.870   -1116.965
sum_light           541.9233     24.030     22.552      0.000     494.818     589.029
sum_worst          1203.0795     61.524     19.555      0.000    1082.477    1323.682
sum_amount_applied    0.0022      0.000     17.093      0.000       0.002       0.002
tr_young farmer     252.1394     36.307      6.945      0.000     180.968     323.311
tr_selected_risk   -203.4869     86.224     -2.360      0.018    -372.507     -34.466
tr_cross_compliance 757.8659     15.367     49.318      0.000     727.743     787.989
tr_number_parcels    17.9997      1.795     10.027      0.000      14.481      21.519
==============================================================================
Omnibus:                    14899.661   Durbin-Watson:                   1.995
Prob(Omnibus):                  0.000   Jarque-Bera (JB):         46814688.607
Skew:                          13.314   Prob(JB):                         0.00
Kurtosis:                     374.347   Cond. No.                     9.08e+05
==============================================================================
```

להלן פלט הרגרסיה שהתקבל:
ניתן לראות כי הרגרסיה מובהקת מכיוון שה-P.value של כל המשתנים המסבירים נמוך מרמת המובהקות 5%.
יתרה מכך, אחוז השונות המוסברת R-square=0.387, מציין כי קיים קשר סטטיסטי בינוני בין הנתונים.

<u>מסקנות הנובעות מחלק ממקדמי הרגרסיה ומן הניתוח:</u>

1. עבור כל יחידה נוספת בפרמטר קנס "קל" סכום הקנס הכולל יגדל בכ-541 יורו. לעומת זאת עבור כל יחידה נוספת בפרמטר קנס "חמור" סכום הקנס הכולל צפוי לגדול בכ-45% יותר מאשר יחידה נוספת בפרמטר קנס "קל". דבר אשר עולה בקנה אחד עם ההנחה שכמות הסיווגים לקנסות שנחשבים ל"חמורים", מגדילה את הקנס באופן משמעותי יותר.
2. עבור מועמדים אשר הוגדרו כצעירים בבקשותיהם סכום הקנס יהיה גבוה בכ-252 יורו יותר מאשר מועמדים מבוגרים. בשילוב המסקנה הקודמת כי קיימת למידה בהליך שליחת הבקשה אצל מועמדים שנקנסו ניתן להניח כי כאשר הצעירים משלמים יותר כך הם לומדים לציית לכללי שליחת הבקשה וככל שהם מתבגרים כך פוחתת הסבירות להיקנס.
3. כל בקשה שעבורה המועמד לא ציית לכללי ההגשה תגדיל את הקנס הכולל בכ-757 יורו.
4. על כל חלקת אדמה נוספת יתווסף לכנס הכולל תוספת של כ-18 יורו.
5. עבור כל יורו שמועמד דורש בסכום הסובסידיה הוא צפוי להיקנס בכ-0.0022 יורו.

<u>משוואת הרגרסיה המתקבלת הינה :</u>

$$y = -1233.92 + 541.92 \cdot sumLight + 1203.08 \cdot sumWorst + 0.0022 \cdot sumAmountApplied + 252.14 \cdot trYoungFarmer - 203.49 \cdot trSelectedRisk + 757.87 \cdot trCrossCompliance + 18 \cdot trNumberParcels$$

לסיכום, ניתן לראות כי הרגרסיה חוזה את גודל הקנס בצורה טובה והמשתנים המסבירים משקפים בצורה אמינה את הקנס הכולל אשר יקבל מועמד .

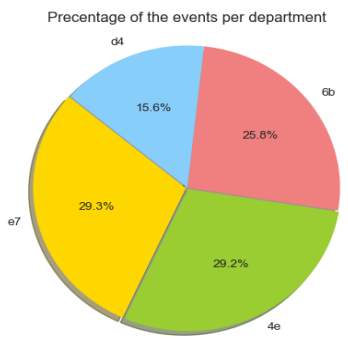
3. <u>עומס על מחלקות ומשאבים</u>

<u>העמודות שנכללות בשאלה:</u> tr_department, org:resource,eventid, tr_year
בחרנו להתמקד בעמודות הנ"ל ממספר סיבות:

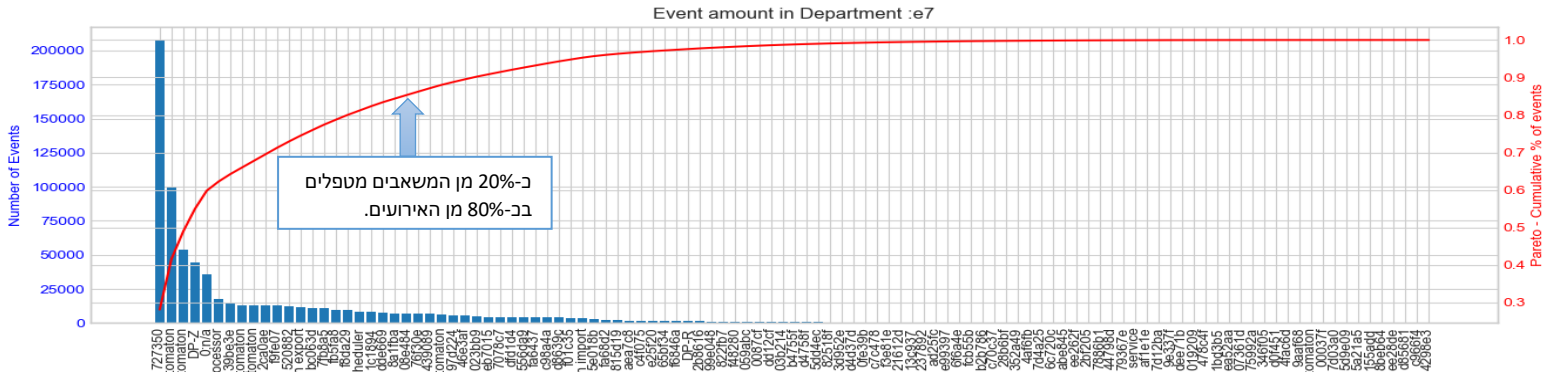
1. בעמודות tr_department-בחרנו כיוון שרצינו לבצע ניתוח נקודתי עבור כל מחלקה.
2. עמודת org:resource-מייצגת את המשאב שה-event עושה בו שימוש.
3. עמודת tr_year- אשר מייצגת את השנה הנוכחית בה הגיש המועמד את בקשתו.

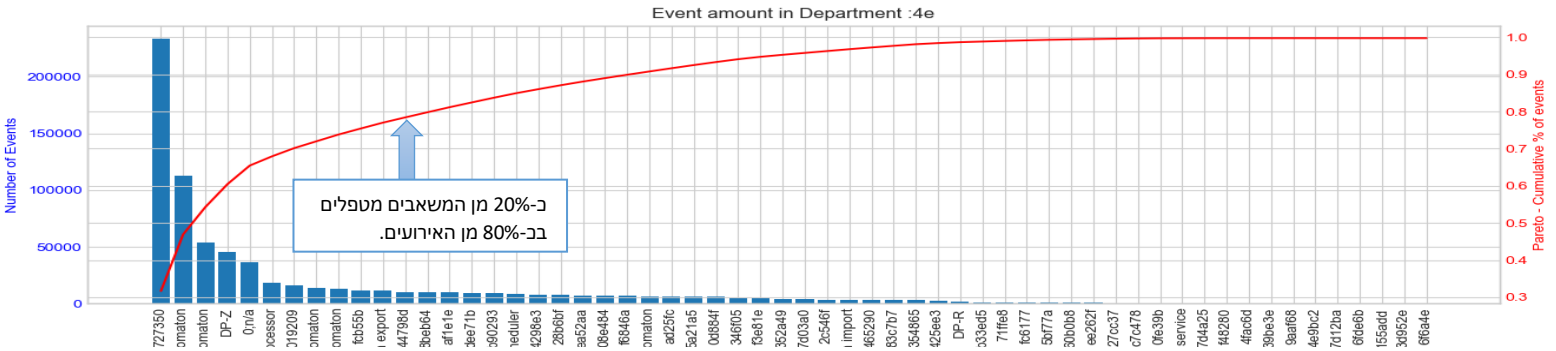
תיאור הניתוח:

- שלפנו מתוך מאגר המידע את העמודות הרלוונטיות לשאלה זו.
- תחילה קיבצנו לפי סוג המחלקה, לאחר מכן, לפי סוג המשאב ולבסוף לפי השנה כאשר לכל קבוצה שכזו ספרנו את מס' האירועים שבה.
- עבור כל מחלקה ספרנו את מס' האירועים המטופלים על ידה, על מנת לחשב מי המחלקה העמוסה ביותר (המטפלת במרבית אירועים).
- יצרנו תרשים פאי הממחיש את התפלגות העומסים על המחלקות השונות: מן התרשים ניתן לראות כי המחלקות העמוסות ביותר הינן מחלקת 4e ו-e7 המטפלות ב29.2% וב-29.3% מכלל האירועים במאגר בהתאמה.



- במסגרת הניתוח בחרנו להתמקד במחלקות אלו ובדקנו מי הם המשאבים העמוסים ביותר ומהו אחוז האירועים שמשויכים לאותו משאב מבין כלל האירועים.
- שני התרשימים הנ"ל מציינים גרפי פארטו להצגת נתונים המאורגנים לפי סדר חשיבות המשאב.





שימוש מקובל בגרף כזה הוא הצגה, בסדר יורד לפי חומרתם, של הגורמים לבעיה, במקרה שלנו- עומס המחלקה הנקבע ע"י שכיחות האירועים בהם מטפלת המחלקה. כך קל לזהות את המשאבים הבולטים ביותר היוצרים את צוואר הבקבוק.

הציר האנכי השמאלי של הגרף מציין את תדירות הופעת המשאב באירועים. הציר האנכי הימני של התרשים מציין את האחוז המצטבר של האירועים המטופלים ע"י המשאב מתוך סך כל האירועים.

<u>מסקנות העולות מן גרפי הפארטו:</u>

תרשים פארטו הוא כלי סטטיסטי שמציג בצורה גרפית את חוק פארטו, ומאפשר לנתח מקרים שמתנהגים בהתאם לחוק זה.

זיהינו כי חוק פארטו אכן מתקיים:

בגרף העליון 22 מבין 109 המשאבים הראשונים, מציינים כ-20 אחוז מהמשאבים במחלקה ומטפלים בקירוב ב-85% מן האירועים השונים במחלקה.

בגרף התחתון 12 מבין 58 המשאבים הראשונים, מציינים כ-20 אחוז מהמשאבים במחלקה ומטפלים בקירוב ב-78% מן האירועים השונים במחלקה.

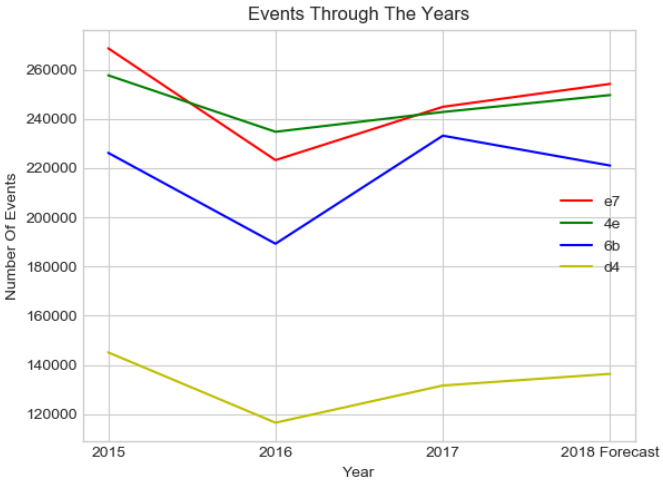
לאחר הסתכלות בגרף עולה באופן מובהק כי 5 המשאבים הראשונים בשני המחלקות זהים והם מטפלים בכמות משמעותית גבוהה יותר ביחס לכל שאר המשאבים.

מבדיקה שערכנו עלה כי :

במחלקת e7: 5 המשאבים הראשונים הינם 4.6% מכלל המשאבים במחלקה ומטפלים ב-59.81% מכלל האירועים.

במחלקת 4e: 5 המשאבים הראשונים הינם 8.6% מכלל המשאבים במחלקה ומטפלים ב-65.6% מכלל האירועים.

לכן, ניתן להמליץ לארגון למנוע את הריכוזיות במשאבים אלו ולפזר את האירועים באופן שוויוני יותר על פני כל המשאבים הקיימים במחלקות או לספק משאבים תחליפיים למשאבים אלו ובכך למזער את העומס ולטפל בצוואר הבקבוק שנוצר עבורם.



<u>החלקה אקספוננציאלית-</u>

ניתוח נוסף שהתבצע היה לספק את תחזית כמות האירועים שיתקבלו לשנת 2018 בהתבסס על מדגם האירועים בשנים 2015-2017.

ביצענו קיבוץ לפי סוג המחלקה ולאחר מכן לפי השנה שבה חל כל אירוע. לכל קבוצה כזו ביצענו ספירה של מס' האירועים שהתקבלו באותה שנה.

בכדי לחזות את כמות האירועים לכל מחלקה בשנה הבאה העברנו את כמויות האירועים המחולקים לפי כל שנה עבור כל מחלקה לפונקציה המבצעת החלקה אקספוננציאלית במשקל של alpha=0.25, משקל זה הינו המשקל שניתן לתצפיות העבר.

<u>מסקנות הנובעות מן ההחלקה:</u>

שתי המחלקות שהוגדרו כעמוסות לעיל, ממשיכות בשנה הבאה להיות העמוסות ביותר וכמות האירועים בהן עשויה להיות דומה.

מחלקת 6b צפויה להיות במגמת ירידה של כ-160 אלף אירועים לעומתה מחלקת d4 עדיין במגמת עליה של כ-50 אלף אירועים אך עדיין צפויה להיות המחלקה בעלת העומס המינימלי.

נמליץ לארגון לבצע חלוקה שוויונית יותר של כמות האירועים על פני המחלקות השונות ובכך לאזן את מצב העומסים בין המחלקות.

מסקנות

בעבודת מחקר זו ניתחנו מאגר המכיל בקשות עבור מתן סובסידיה לחקלאים באירופה, חקרנו את תהליך הגשת הבקשה ואת ההשלכות אשר נלוות לתהליך זה. כחלק מן המחקר שבוצע ניתחנו את הגורמים המשפיעים על דחיה או אישור הבקשות של המועמדים, הערכנו את גודל הקנס העלול להתקבל בעקבות חריגות בשליחת הבקשה וניתוח עומסים שונים בין המחלקות ובתוכן. התוצאות שלנו מציינות נקודות השערה מעניינות העשויות לספק תובנות בנוגע לתהליך העסקי ומציגות תמונה כוללת על שלבים שונים הנכללים בתהליך זה.

מחקר זה עשוי לשמש את האיחוד האירופי כדרכים לייעול ושיפור הליך הגשת הבקשה על ידי החוואים במדינות השונות באירופה ואת יכולת הארגון באופן טיפול הבקשות.

במהלך האנליזה שבדקה את הקשר בין פרמטרים שונים לסיכויי הבקשה להידחות, מצאנו מודל חיזוי המסווג באופן מיטבי האם בקשה תאושר או תידחה. מודל זה עשוי להקל על המועמד בשליחת הבקשה בכך שתציג בפני המועמד מבעוד מועד על סמך ההיסטוריה האישית שלו את סיכויו לקבלת אישור ותשלום ישיר עבור בקשתו וכמו כן לייעל את הליך הבקשה מצד הארגון בכך שיהיה ניתן להשתמש במודל כבדיקה ראשונית עבור סינון בקשות, דבר שיסייע בקיצור משך הבקשה, יפחית את השלבים עבורה ואת זמן הטיפול שמיועד לה על ידי המשאבים השונים.

מעבר לכך, גילינו שבקשות המוכלות במסמך ה- Parcel Document (מסמך השייך לתהליך העסקי) יקבלו קנס בהסתברות גבוהה. מכאן, הסקנו שיש להתייחס לקנסות השונים ולנתח את הגורמים שיובילו לעלייה בסכום הקנס שמועמד עלול לקבל. בעזרת מודל רגרסיה ליניארית גילינו שקיימת השפעה של גורמים שונים לסכום הקנס ואת התוספת היחסית שלהם לגודלו. כמו כן, זיהינו כי קיים תהליך של למידה בקרב החקלאים בעקבות הקנסות שניתנו להם.

יתרה מזאת, התברר לנו שמחלקות מסוימות בארגון עמוסות יותר בהשוואה לאחרות , לא זאת בלבד, אלא שישנם משאבים בתוך המחלקה שעמוסים באופן לא שוויוני. בנוסף, ביצענו תחזית לכמות הבקשות שיתקבלו בשנת 2018 עבור כל מחלקה וראינו כי קיימות עבורן מגמות שונות.

המלצתנו לארגון היא לבחון את המצב הקיים בו ישנו אי שוויון מבחינת כמויות האירועים ,הפתרון לכך הוא לשלב משאבים תחליפיים עבור המשאבים העמוסים ביותר במחלקות ולפזר באופן מתון את כלל האירועים המטופלים בין המחלקות ובתוכן.