**Titanic: Machine Learning from Disaster**

**לירן זלצברג 305631731**

**אדווה חוטה 203436654**

1. **תיאור הבעיה:**
   1. **הצגת התחרות:**הטביעה של אונית הטיטניק הוא אחד האסונות הידועים ביותר לשמצה בהיסטוריה.   
      ב -15 באפריל 1912, במהלך מסע הבכורה שלה, טיטניק שקעה לאחר התנגשות עם קרחון, ונהרגו 1502 איש מתוך 2224 נוסעים. אחת הסיבות לאובדן חיים בסדר גודל זה היא שלא היו מספיק סירות הצלה לנוסעים ולצוות. למרות שמדובר בעניין של מזל, היו קבוצות מסוימות של אנשים ששרדו יותר מאחרים : כמו נשים, ילדים ומעמד עליון.
   2. **מה רוצים לגלות?**

בתחרות זו נתבקשנו לנתח אילו סוגים של אנשים היו עשויים לשרוד. ובפרט, נתבקשנו ליישם את  
הכלים של ML כדי לחזות אילו נוסעים שרדו את הטרגדיה.

* 1. **על סמך מה?**

התחרות מספקת נתונים על הנוסעים שבעזרתם יש לבנות מודל חיזוי עבור גורל הנוסעים.

* 1. **סוג הבעיה איתה אנו מתמודדים:**

התחום ב ML –בו אנו עוסקים הוא Supervised learning - טכניקה ללמידת פונקציה על בסיס סט אימון – למידה עם "מורה". טכניקה זו עוסקת בבעיה שבה המטרה היא להצמיד label לכל דוגמא.   
כל דוגמא מתוארת על ידי מידע רלוונטי. לאלגוריתם הלומד יש גישה לקבוצה של דוגמאות מתויגות.  
כאשר מטרתו היא ללמוד כלל המקשר בין תיאור הדוגמא לתווית המתאימה לה. הכלל שלמד האלגוריתם יופעל על דוגמאות חדשות, שהתווית שלהן לא ידועה, ויחזה מהי התווית הנכונה לכל דוגמא. כלל הוא מוצלח אם הוא חוזה את התווית הנכונה באחוז גבוה של הדוגמאות החדשות.

אנו מתמודדים עם בעיה מסוג Binary classification שבה יש לסווג סט של נתונים לשני קבוצות על פי כלל בסיס שלפיו הסיווג יעבוד.

* 1. **Data :**

ה – Data חולק לשתי קבוצות:

training set (train.csv) – 891 רשומות של נוסעים בהם מופיעים ה-labels המקוריים, עמודת Survived (1\0) . בעזרת Data זה עלינו לאמן את המודל אשר בעזרתו נחזה את התוצאות.



test set (test.csv) – 418 רשומות של נוסעים אשר אינם מופיעים בסט ה-train ולא קיימם עבורם labels ( (Survivedובעזרתם נבדוק את טיב המודל שלנו ונחזה עבור כל נוסע מסט זה האם הוא שרד או לא.



* *Survival* – 1 שרד 0 מת.
* *Pclass* – מחלקה בה הנוסע שהה בספינה.
* *Sex* – מין הנוסע.
* *Age* – גיל.
* *Sibsp* – כמות קרובי משפחה\בני זוג על הספינה.
* *Parch* – מספר ההורים\ילדים של הנוסע על הספינה.
* *Ticket* – מספר כרטיס.
* *Fare* – מחיר ששילם נוסע עבור הכרטיס.
* *Cabin* – מספר תא.
* *embarked* - נמל ממנו יצא הנוסע

((C = Cherbourg, Q = Queenstown, S = Southampton

* 1. **חוסרים בסט האימון :** סך הכל 891 רשומות בסט האימון.

|  |  |  |
| --- | --- | --- |
| **Percentage** | **Missing** | **Feature** |
| 77% | 687 | Cabin |
| 19% | 177 | age |
| 0.2% | 2 | Embarked |

* 1. **מדד לביצוע :**

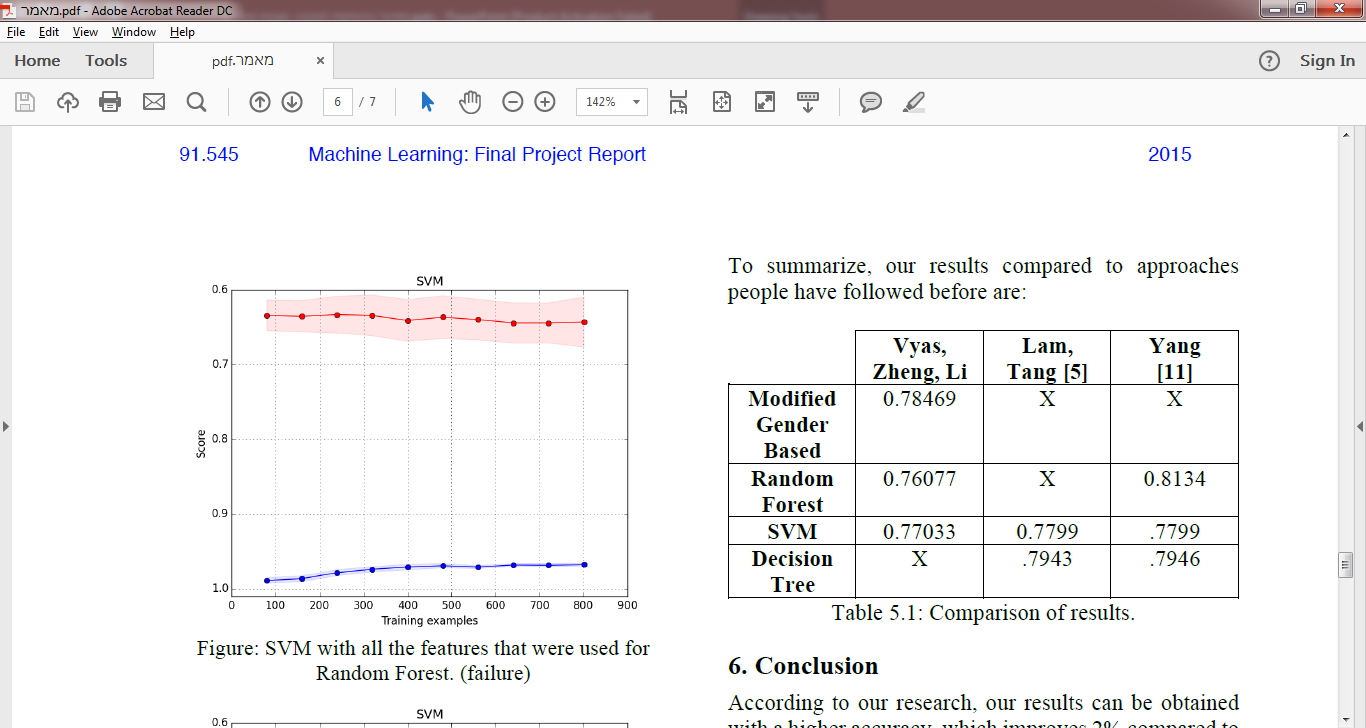
מגישים לתחרות את טבלת תוצאות ה test . הטבלה מכילה עבור כל passengerId את הערך survived שהמודל שלנו חזה עבורו. המדד לביצוע הוא אחוז הדיוק של המודל עבור נתונים אלו.

1. **סקירת ספרות:**

המאמר [Titanic - Machine Learning From Disaster](mailto:10.1109@IDEAS.2004.1319778.pdf)מציג את בעיית הטיטאניק באופן זהה לצורה בה הוצג בתחרות. הבעיה איתה המאמר מתמודד היא גם חיזוי תוצאות עבור כל נוסע בספינה ומוצגות השיטות שבה השתמשו מחברי המאמר, ועקרונות מנחים עליהם התבססו. המאמר מציג את התוצאות הסופיות בהשוואה לחוקרים אחרים אשר התמודדו עם הבעיה ומפרט על השוני בחיזוי וההתייחסות לנתונים.

**2.1 הנחות עליהם התבססו החוקרים במאמר :**

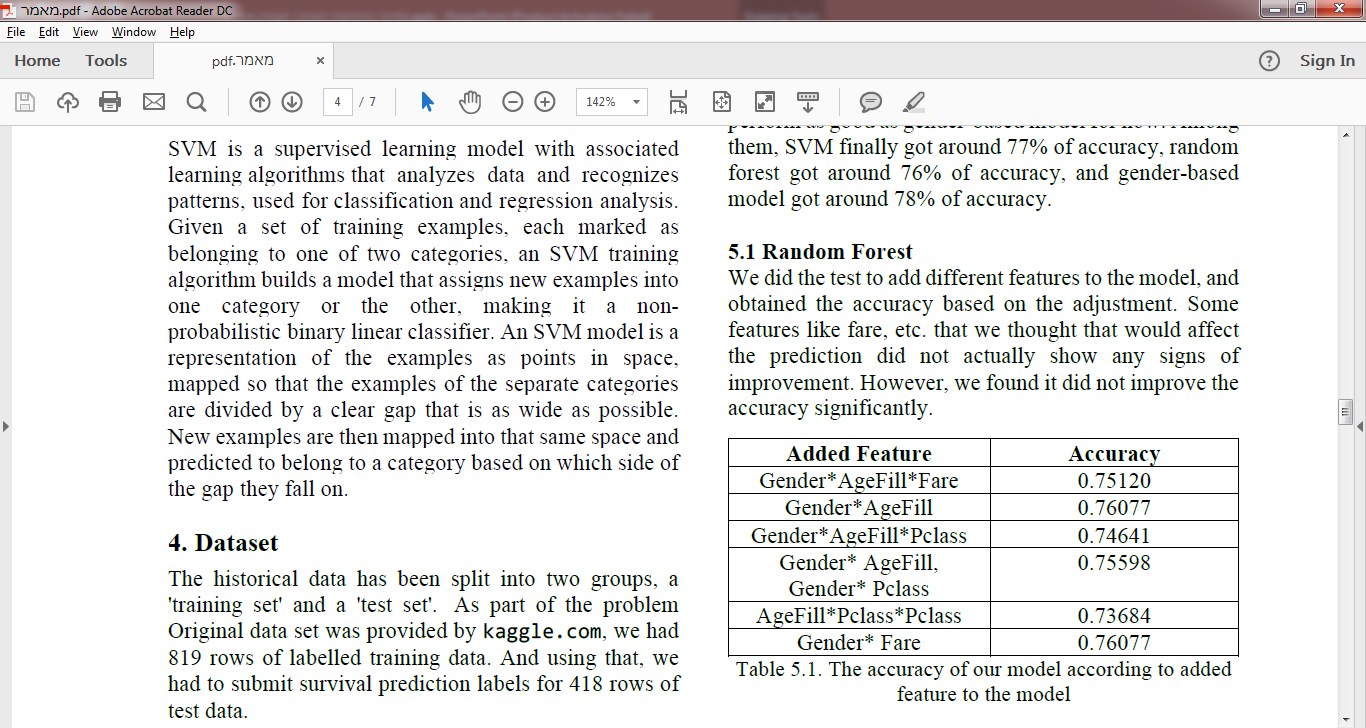
* ילדים, נשים ואנשי המעמד הגבוה בעלי סיכויי הישרדות גבוהים יותר.
* כמות גבוהה יותר של features לא בהכרח נותנת תוצאות טובות יותר.
* -"women and children first"כמו בכל אסון עקרון זה מלווה את המחקר כהנחת יסוד.
* ישנה הבדלה גם בין המחלקות בספינה לדוגמא: נשים מהמחלקה ה -3 שרדו יותר מגברים מהמחלקה ה -1.
* class הוא feature חלש יותר מ- sex או age.
* נעשתה השוואה מול חוקרים אחרים שהתמודדו עם אותה בעיה וניסו לשפר את רמת הדיוק על ידי בחירת features שונה משלהם.
* ה-feature -cabin הושמט מכיוון שהיו חוסר גדול מדי ברשומה זאת.
* בוצעה השלמה של החוסרים ב-age לפי חציון, מכיוון שרוב הנוסעים היו בין הגילאים 20 – 30 והרוב בגיל 22, הם בחרו את 22 כחציון.  
  1. **השיטות בהם השתמשו במאמר:**
* *Random Forest.*
* *SVM - RBF kernel.*
* *Improved Gender Model –* המודל מבוסס על מערך תלת ממדי שתפקידו לנתח סיכוי הישרדות. לאחר שהגיעו למסקנה כי הGender הוא משתנה מכריע ביותר הם בחרו להוסיף שני משתנים נוספים שיהיה להם קשר חזק לגורל הנוסע – fare , pclass שלדעתם הם בעלי קשר גבוה מכיוון שקשורים למעמד ולכסף.
  1. **התוצאות:** החוקרים הגישו את תוצאות המודלים שלהם ל –kaggle והציגו את התוצאות שקיבלו בהשוואה לחוקרים אחרים (טבלה 1א).



טבלה 1א

* 1. **ניתוח התוצאות ומסקנות:**
* *Random Forest:*בוצע ניסיון לבדוק מספר קומבינציות של features לשם שיפור הדיוק אך לא נראה שינוימשמעותי. (טבלה 1ב)
* *SVM:* לאחר שניסו מספר קומבינציות של features, בוצע גם feature scaling ל-fare שהוא מעין סוג של נרמול לfeature ה"סובל" מטווח רחב מדי של ערכים. היה ניסיון לא מוצלח לבצע את השלמת הערכים החסרים עבור age ע"י logistic regressor המבוסס עלEmbarked, Parch, SibSp, Pclass במקום להשתמש בחציון אך הדיוק היה מאוד נמוך. לאחר שביצעו feature reduction נותרו עם pclass, embarked, sex ולהפתעתם הגיעו לתוצאות הכי טובות לאחר שהורידו את שאר הfeatures שביניהם היה age שלא תיארו שיחליש את המודל.

טבלה 1ב

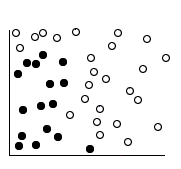


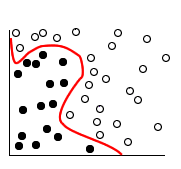
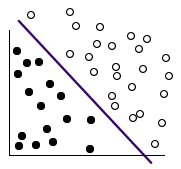
* *Improved Gender Model:* ככל שמחיר הכרטיס היה גבוה יותר עלו סיכוי ההישרדות – נוסעים עשירים. תחיליות לשמות כמו Mr./Mrs./Miss/Dr./Master/Col/Rev ניחנו בסיכויי הישרדות של 50%. היה קשר בין הנמל ממנו יצאו הנוסעים לבין סיכויי ההישרדות. לגודל המשפחה היה קשר לסיכויי ההישרדות - משפחה גדולה יותר הקטינה את סיכויי ההישרדות.

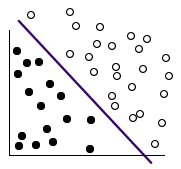
1. **מתודולוגיה בה נעשה שימוש בפרויקט:**

התחום בו אנו עוסקים הוא supervised learning.  
הבעיות בהן עוסקת supervised learning כוללות בין היתר רגרסיה (regression ) ובעיות סיווג (classification).RFM וSVM ו - Logistic regression הן שיטות מסוג supervised learning.

השתמשנו ב-3 שיטות אלו בפרויקט זה :

1. Random forest model (RFM) : מודל המבוסס עצי החלטה שתפקידו לחזות תוצאה מסוימת.  
   עץ החלטה הוא מודל המספק מיפוי בין תצפיות לערכים המתאימים עבורם. עץ ההחלטה יכול לשמש כמודל חיזוי הממפה תצפיות על פריט ויוצר מסקנות על ערך היעד של הפריט.  
   העץ הוא בינארי מלא המורכב מצמתי החלטה שבכל אחד מהם נבדק תנאי מסוים על feature מסוים של התצפיות, ועלים המכילים את ערך החיזוי עבור התצפית המתאימה למסלול שמוביל אליהם בעץ. לבסוף עץ ההחלטה יצור מעין **"**חוק" .  
   כל העצי החלטה יצרו Random Forest - טכניקת הכללה אשר בונה יותר מעץ החלטה אחד לצורך שיפור דיוק הסיווג. מחלקים את סט האימון רנדומלית למספר תתי סטים, לכל תת סט בונים עץ החלטה, מקבלים יער שלם של עצי החלטה, לוקחים דגימה מסוימת מסט האימון ובודקים בכל עץ את הסיווג שנבחר עבורה, לאחר שאוספים את כל הסיווגים בוחרים על פי רוב את הסיווג הסופי.
2. *SVM – :support vector machine SVM*   
   בטכניקה זאת מפרידים בין הנקודות -שמייצגות את דוגמאות האימון - לשתי מחלקות : חיובית ושלילית.  
   המוטיבציה היא ליצור וקטור מפריד בין שתי המחלקות, כאשר הווקטור המיטבי הוא זה שבעל המרחק הגדול ביותר משתי המחלקות. לאחר שסיווגנו את הדוגמאות, כאשר נקבל דוגמא חדשה האלגוריתם ידע לזהות אם היא בתוך הקו שמגדיר את הקבוצה או מחוצה לו.SVM עבור רב ממד – SVM ממפה data למרחב רב ממדי (לפי הfeatures ) כך שנוכל לסווג נקודות המייצגות נתונים (רשומות). האלגוריתם מוצא hyperplane אשר מפריד בין הקטגוריות שנמצאו.



1. logistic *Regression:*  
   מודל רגרסיה בו המשתנה התלוי הוא קטגוריאלי. כלומר, נע בטווח מוגדר מסוים. זוהי טכניקה המתאימה לבעיות מסוג binary classification – בעיות עם תוצאה בינארית. שיטה זו משתמשת בlogistic function– פונקציה המתארת עקום בצורת s . הפונקציה לוקחת ערכים מספריים אמיתיים וממפה אותם לערכים בין 0-1 אבל לא בדיוק. הפונקציה : כאשר ה- value זה הערך המספרי שנרצה להמיר.  
   Logistic Regression לוקחת מספר קלטים בעלי ערכים אמיתיים ועושה "חיזוי" לגבי ההסתברות של הקלט להיות שייך למחלקת הdefault .אם ההסתברות גדולה מ0.5 הקלט שייך למחלקת default אחרת הוא שייך למחלקה השנייה. ה- output נקבע על ידי :

output = b0 + b1\*x1 + b2\*x2

לכן קודם כל נבצע "אלגוריתם לומד" שתפקידו יהיה לגלות את הערכים הטובים ביותר למקדמים בעזרת הtraining Set.. לאחר מכן נבצע את האלגוריתם על הtest ו"נחזה" את התוצאות. ה output מתורגם להסתברות על ידי הפונקציה כאשר value = output .

1. **תוצרי הפרויקט:**
   1. **השלמת החוסרים בסט האימון:**

|  |  |  |  |
| --- | --- | --- | --- |
| **כיצד השלמנו** | **Percentage** | **Missing** | **Feature** |
| השמטנו בגלל חוסר רב מידי | 77% | 687 | Cabin |
| ממוצע גיל :29 | 19% | 177 | age |
| נשאר כמו שהו | 0.2% | 2 | Embarked |

* 1. **השלמת החוסרים בסט האימון:**המודל צריך לעבוד עם משתנים נומריים בלבד לכן השתמשנו בפונקציות שתפקידן להמיר את הערכים הקטגוריאליים – **sex , embarked** לערכים בינאריים.   
     **ticket** - סיווגנו משתנה זה לערכים בינאריים לפי התחילית של מספר הכרטיס. משתנה ללא תחילית יסווג לערך XXX.   
     **family size -**  בעזרת ניתוח שני ה features :   
      *Parch* – מספר קרובי המשפחה / בעל / אישה על הספינה.  
      *sibsp* – מספר ההורים / ילדים של נוסע על הספינה.  
     חישבנו את ה family size וסיווגנו את התוצאות ל-3 קטגוריות :  
      family single- נוסע ללא קרוב משפחה על הספינה.  
      family small- נוסע בעל 2-4 קרובי משפחה על הספינה.  
      family large– נוסע בעל 5 ומעלה קרובי משפחה על הספינה.   
       
     ביצענו ניתוח מקדים כדי לקבל תחושה על הקשר בין משתנים מסוימים לסיכויי ההישרדות.  
     השתמשנו בפונקציה שנקראת plot\_categories פונקציה שהגדרנו שמקבלת את המשתנה הקטגוריאלי ואת הlabel של התוצאה. ויוצרת גרף שמעידה כמה מהקטגוריה הזו שרדו.
  2. **בנינו מודל לכל שיטה עם המשתנים : age , fare , pclass , sex , familySize , embarked , ticket**

|  |  |
| --- | --- |
| **הדיוק** | **השיטה** |
| **0.75119** | RFM |
| **0.61244** | SVM - rbf |
| **0.76555** | SVM- linear |
| **0.66028** | SVM - poly |
| **0.62679** | SVM- sigmoid |
| **0.76076** | Logistic regression |

* 1. **feature selection -** בחירה אוטומטית של ה features שהכי רלוונטיים למודל .  
     השתמשנו בשיטת feature importance - שיטה המבוססת על עצי החלטה בה יוצרים משתנה החלטה מלאכותי שתפקידו לסווג לפי החשיבות של ה-features . החשיבות נקבעת על ידי מספר צמתי ההחלטה בו ה feature משתתף. השיטה מיושמת רק על מודל ה- RFM שמבוסס עצי החלטה בעצמו. מתוצאות הניתוח ראינו כי המשתנים שעלו כחשובים ביותר הם : age , fare , sex .השתמשנו בתוצאות הניתוח של ה-feature importance גם עבור המודלים האחרים.

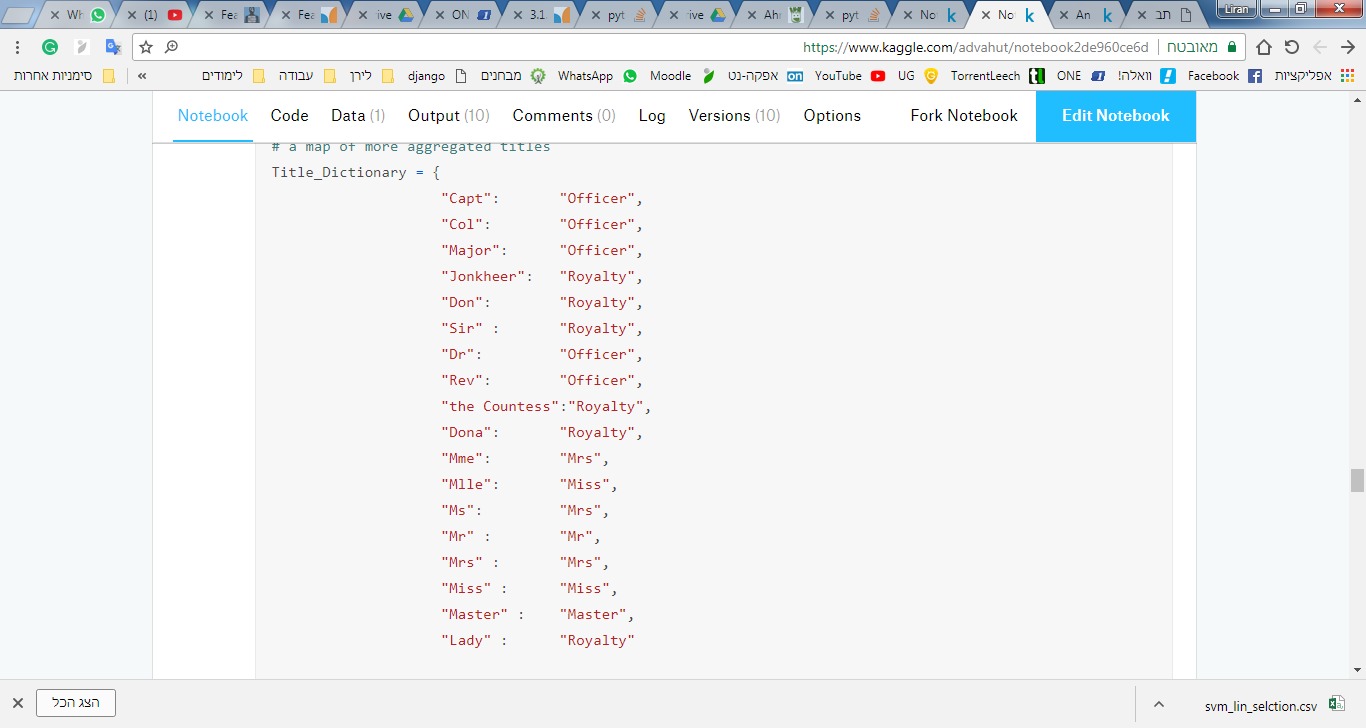
|  |  |
| --- | --- |
| **הדיוק** | **השיטה** |
| **0.71770** | **RFM** |
| **0.76555** | **SVM- linear** |
| **0.75598** | **Logistic regression** |

ביצענו פעולה אחרת בתחום feature selection שהיא selectFromModel אשר מסננת feature לפי חסם מסוים ומאמנת את המכונה עם מודל ספציפי.

|  |  |
| --- | --- |
| **הדיוק** | **השיטה** |
| **0.77033** | **SVM- linear** |

מכיוון שאחד היתרונות הבולטים של RFM הוא היכולת לבצע את המודל גם עם משתנים בעלי חוסרים, הרצנו את מודל ה RFM כדי להבין האם החוסר העצום ב cabin feature באמת ישפיע על המודל לפי ההנחה.

|  |  |
| --- | --- |
| **הדיוק** | **השיטה** |
| **0.79425** | RFM |

****הוספנו משתנה שנקרא title - משתנה הבנוי ממילון אשר מכיל מפתח וערך.   
המפתח הוא התחילית של השם והערך מסמל את מעמדו.   
הביצועים לאחר ההוספה עם כל הfeatures שקיימים בתחרות :

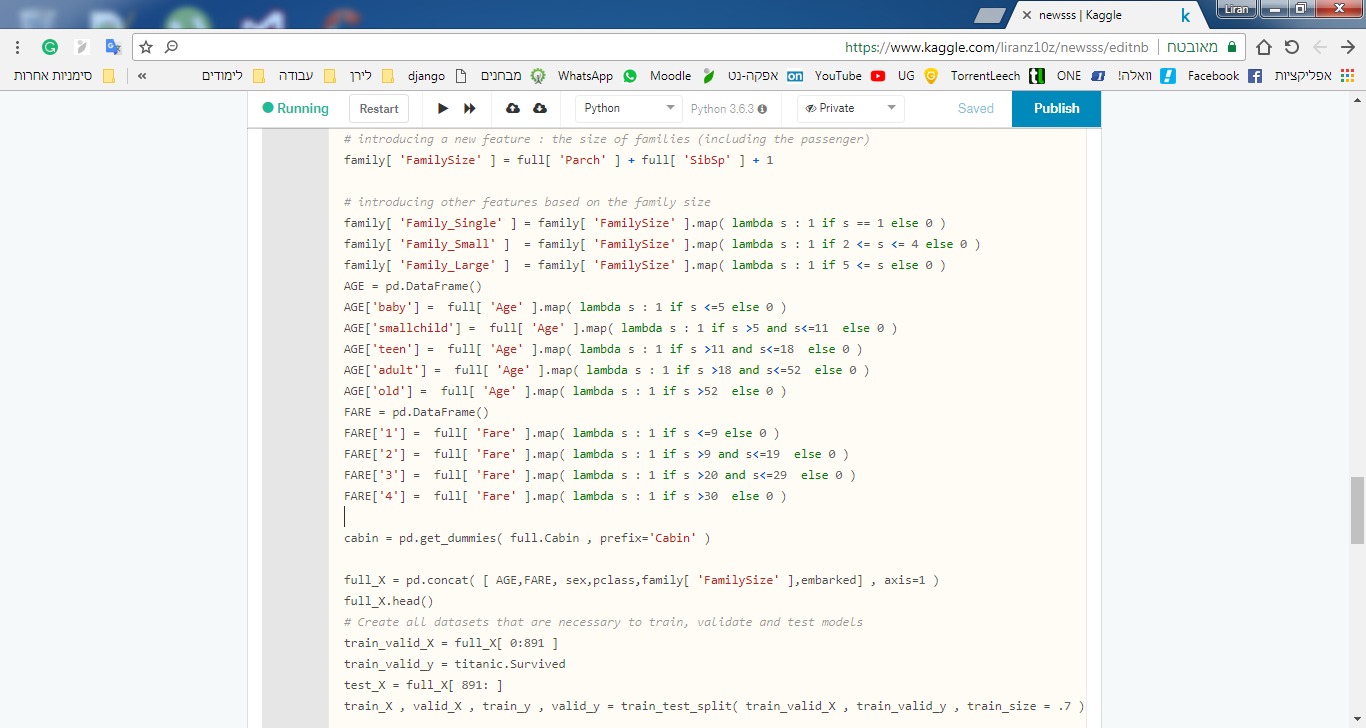
|  |  |
| --- | --- |
| **הדיוק** | **השיטה** |
| **0.74** | RFM |
| **0.76** | SVM- linear |
| **0.75598** | Logistic regression |

ביצענו **K fold cross validation –** ניתוח סטטיסטי שנותן מדד ליציבות המודל.

\*צריך להוסיף עוד K.

נשים לב לסטיית התקן . כאשר הסטייה גדולה יותר בין K שונים כך המודל פחות יציב. נראה כי בהשוואה לשאר המודלים RFM הוא בעל הסטייה הקטנה ביותר ולכן הוא נחשב למודל היציב ביותר מבין המודלים השונים שבדקנו.

|  |  |  |  |
| --- | --- | --- | --- |
| **המדד לפי K** | **השיטה** | | |
| **K=10** | **K=5** | **K=2** |  | |
| 0.79 | 0.79 | 0.78 | RFM | |
| 0.81 | 0.81 | 0.78 | SVM- linear | |
| 0.80 | 0.80 | 0.77 | Logistic regression | |

**ניסיונות נוספים לשיפור הדיוק :**   
באמצעות חלוקה של features נוספים לקטגוריות קבועות מראש –

|  |  |  |
| --- | --- | --- |
| **לאחר השינוי** | **הדיוק לפני השינוי** | **השיטה** |
| 0.73205 | 0.74 | RFM |
| 0.76076 | 0.76 | SVM- linear |
| 0.77033 | 0.75598 | Logistic regression |

1. **מסקנות:**

* קיימת חשיבות גדולה לשלב המקדים של הכנת ה – data , השלמת החוסרים וחלוקה נכונה לקטגוריות של features בעלי ערך נומרי.
* הבחנו בכך שהיו מודלים מסוימים (logistic regression ) שהתוצאות שלהם השתפרו עקב עבודה מקדימה זו.  
  במודל logistic regression יש חשיבות לfeatures שנעים בטווח קבוע מראש.
* לאחר ביצוע הfeature selection אשר תוצאותיו היו fare , age , sex , ראינו כי ביצוע המודלים אך ורק על משתנים אלו לא בהכרח משפרים את דיוק המודל . אך זה שיפר את מהירות החישוב בצורה מובהקת.
* לעומת זאת כאשר ביצענו את פעולת select from model על ה- SVM חל שיפור.
* רצינו לבדוק במהלך העבודה האם החוסר בcabin ישפיע על הRFM. הופתענו לגלות שזהו המודל הטוב ביותר שביצענו עד כה.
* לפי מדד ה- K fold cross validation ראינו כי מודל SVM-linear הוא המדויק ביותר.

**מסקנות לעתיד :** קיימים מספר דרכים שאנו חושבים שעלולות לשפר את המודל :

* שיטות יצירתיות ליצירת features חדשים על ידי ניתוח מעמיק וגילוי קשרים חדשים בין הfeatures הקיימים.
* לנסות מודלים שלא מימשנו בסמינר למשל Naïve Bayes שביצעו חוקרי המאמר שקיבלנו השראה ממנו.

1. **ביבליוגרפיה:**

* **Titanic - Machine Learning From Disaster**Kunal Vyas, Zeshi Zheng, Lin Li  
  2015
* **HOW THE RANDOM FOREST ALGORITHM WORKS IN MACHINE LEARNING**May 22, 2017 [Saimadhu Polamuri](http://dataaspirant.com/author/saimadhu/)  
  http://dataaspirant.com
* **https://en.wikipedia.org/wiki/Random\_forest**
* <https://en.wikipedia.org/wiki/Support_vector_machine>
* <https://www.youtube.com/watch?v=1NxnPkZM9bc>
* <https://en.wikipedia.org/wiki/Logistic_regression>
* <https://machinelearningmastery.com/logistic-regression-for-machine-learning/>