

**אוניברסיטת תל-אביב**

**הפקולטה לניהול ע"ש קולר**

**החוג לניהול**

**פרויקט מסכם**

**קורס** " **כריית מידע עסקי**"

**מרצה:**

**ד"ר תומר גבע**

**מגישים:**

**אלון דוידוף 204478531**

**אושרת ביטון 304986524**

**תאריך ההגשה: 18.1.2018**

**דוח מנהלים**

הבעיה העסקית שעומדת לפנינו היא בחינת שימוש בטכנולוגיות כריית מידע לצורך שיפור תהליכים עסקיים. מתוך הבעיה הרשומה לעיל, אנו הגדרנו כי הבעיה העסקית היא מקסום רווחי החברה על ידי פניה ללקוחות המתאימים. מטרתנו היא למצוא מודל חיזוי שידע לנבא בצורה הטובה ביותר מיהם הלקוחות תניב רווח.

בהתאם להגדרת הבעיה העסקית הרשומה לעיל, מדד הביצועים הרלוונטי הוא רווחיות החברה. את רווחיות החברה מדדנו באופן הבא: החברה יכולה לבדוק את הרווחיות שלה מפניה לכל הלקוחות ואת הרווחיות מאי פניה לכל הלקוחות. כאשר:

170\*TP – הלקוחות שחזינו שיקנו וצדקנו – ולכן יש לנו מכל לקוח רווח של 170 דולר.   
32\*FP – הלקוחות שחזינו שיקנו וטעינו – ולכן יש לנו הפסד של 32 דולר מהפניה אליהם.   
  
טרם התחלת העבודה בדקנו את תקינות הנתונים - וידאנו שאין ערכים חסרים ואכן לא היו כאלה ולא ביצענו מניפולציות על המידע, פרט למשתנה המייצג את הביקורות של המשתמשים, אשר יפורט עליו בהמשך.

**בחירת סף הסיווג:**

בדקנו מהו סף הסיווג הרלוונטי עבורנו בהינתן השקול הכלכלי , ובחרנו להשתמש ב Expected Value Framework . על מנת שהמודל יחזה את רווחיות, EV צריכה להיות גדולה מ – 0. קבענו שההסתברות המינימלית להצלחה על מנת שנרצה לטרגט את הלקוח, כלומר הלקוחות שנפנה אליהם בהינתן רווח נקי של 170$ והפסד מפניה ללקוח שלא משכיר רכב של 32$ , הם אותן לקוחות שההסתברות שהם יקנו גבוהה מ 0.1584 . [ הנוסחה : ].

**החלטת המשתנים האינפורמטיביים:**

בדקנו מה מתוך ה 21 משתנים הם אינפורמטיביים על ידי הרצת Random Forest – ומתוכו כתבנו את הפקודה " Importance" על מנת לראות מהי מידת האינפורמטיביות של כל משתנה. ראינו כי ישנם 3 משתנים שקיבלו את הערך הנמוך ביותר (המשתנים הם Status Platinum, Status Gold, Call center flag ,כאשר SP קיבל את המדד הכי קטן, מתחת ל -1 , אחריו SG , גם מתחת ל-1, ולבסוף CCF מעט מעל 1). שאר המשתנים נמצאות כאינפורמטיביים וקיבלו מדד גבוה יחסית. לכן, כאשר הרצנו את המודלים השונים, הרצנו בכל פעם את המודלים ללא כל אחד מן המשתנים הרשומים לעיל, וללא קומבינציות ביניהם בהתאם למידת האינפורמטיביות.

**טיפול במשתנה Reviews** :

המשתנה Reviews הוא משתנה שמקבל ערכים של 0,1, 1- בהתאם לחוות הדעת של המשתמשים כאשר 0 אומר אין חוות דעת לגבי אותו לקוח, 1 אומר הלקוח נתן חוות דעת חיובית ו – 1- אומר הלקוח נתן חוות דעת שלילית. גם כאן, השתמשנו ב Random Forest על מנת לקבוע האם המשתנה Reviews הוא חיובי או שלילי . לאחר מכן השתמשנו ב Excel ובפונקציית השונות שבו, ביניהן ה Vlookup על מנת להכניס את משתנה ה Reviews אל תוך נתוני ה Training .

המדד שלפיו בדקנו את מודל ה Random Forest הוא מדד ה- Accuracy כאשר במודל הסופי הוא 82.1.

**הפעלת שיטות רלוונטיות:**

הבעיה שעומדת לפנינו היא - בהינתן כל המשתנים הקיימים לרשותנו ( X) מה ההסתברות שלקוח ישכיר או לא ישכיר רכב (Y). כיוון והבעיה שעומדת לפנינו הינה בעיה מונחית (משתנה המטרה מוגדר ויש לנו נתונים עליו) החלטנו לבחור בשיטות מידול של סיווג ורגרסיה – ולכן שיטות המידול שאותן הרצנו הן : Random forest, Logistic Regression , Naïve Bayes והרצנו גם מודל פשוט – חזינו שאף לקוח לא ישכיר רכב. בחרנו להריץ מודל זה כמודל Benchmark , מודל חיזוי פשוט שיעזור לנו להבין איך הביצועים של שאר המודלים שלנו ביחס אליו. זהו בעצם האלגוריתם הבסיסי והנאיבי ביותר שבחרנו להריץ.

במהלך הרצת המודלים השונים הפעלנו מספר שיטות מידול – תחילה, הרצנו כל מודל בפני עצמו. פעם הראשונה הרצנו את כלל המודלים ללא הכנסת משתנה ה- Reviews ובפעם השנייה את כלל המודלים עם הוספת המשתנה. לאחר ההרצה הפשוטה של כל מודל בפני עצמו, עבור כל מודל השמטנו משתנים / שינינו את אופן הפעולה בהתאם לתוצאות שקיבלנו . כך למשל את מודל ה Random forest הרצנו פעם אחת עם 100 עצים, בפעם השנייה עם 500 עצים, ובפעם השלישית עם 500 עצים ובלי המשתנה SP . ניתן לראות כי הן ללא הוספת משתנה ה Reviews והן עם הוספת המשתנה, הרווח גדל ככל שמספר העצים גדל. כאשר הרצנו את Random forest עם 500 עצים ללא המשתנה SP וללא ה Reviews הרווח קטן, בעוד שבמודל עם ה Reviews הרווח דווקא גדל , הסבר אפשרי לכך הוא שמשתנה ה SP כנראה מוסבר טוב על ידי משתנה ה Reviews ולכן דווקא הוספה שלו משפרת את חיזוי המודל ולא גורעת. בעבור מודל הרגרסיה הלוגיסטית הרצנו את הקומבינציות של בלי SG , בלי SP , בלי SP&SG ובלי SG&SP&CCF. כאשר הוספנו את המשתנה CCF ראינו כי רווחיות המודל דווקא קטנה ולכן כנראה שמשתנה זה הוא אינפורמטיבי. השינוי היה כאשר הרצנו את המודל Naïve Bayes . תחילה, הרצנו את המודל עם כל המשתנים וקיבלנו רווח שלילי. לאחר שווידאנו שכל הקוד תקין, עלתה בדעתנו המחשבה שאולי יש במודל בעיית Over fitting ראינו כי המודל נתן הסתברות גבוהה מאוד להיות "True" ושרווח המודל יוצא שלילי ,ולכן התחלנו להריץ את המודל מספר פעמים כאשר בכל פעם השמטנו משתנה אחד מתוך ה – 21 שיש לנו. במהלך ההרצות קיבלנו כי השמטה של המשתנה KM\_L\_Y1 אומנם מקטינה את הרווח השלילי המתקבל אך לא מסירה אותו לגמרי. כאשר הסרנו במקביל אליו את משתנה ה SP קיבלנו חזרה רווח חיובי במודל. באופן זה, הרצנו את אלגוריתם Naïve Bayes ללא המשתנים שיצרו את ה Over fitting כאשר בכל פעם השמטנו משתנים אחרים שראינו כי הם פחות אינפורמטיביים – פירוט ניתן לראות בנספח א'. (יש לציין כי לא כלל ההרצות שתאורו לעיל נמצאות בנספחים למעט ההסרה של אותו משתנה והמשתנים האחרים שמצאנו כי הם לא אינפורמטיביים מספיק). באופן זה קיבלנו 31 מודלים אפשריים. בנספח א' ניתן לראות כי כל התוצאות עם המשתנה Reviews היו טובות יותר מאשר התוצאות ללא המשתנה.

**אופן הטיפול בנתונים** :

את כלל הנתונים שלנו חילקנו באופן הבא: 30% הולך לנתוני ה Test , ואת ה- 70% הנותרים חילקנו ל- 70% לנותני Training ו-30% לנתוני הוולידציה. המוטיבציה לשיטת הפעולה הנ"ל – מקובל לחלק את הנתונים ל Test ,Training, & Validation כאשר יש לנו כמות רשומות גדולה מאוד , וכאשר יש לנו הרצה של המון מודלים. אומנם כמות הרשומות שיש לנו היא לא גדולה, אך כמות המודלים שהרצנו גבוהה מאוד ( 31 מודלים) ולכן בחרנו כן לחלק את הנתונים בצורה זו.

**תוצאות :**

לאחר כל ההרצות הנ"ל קיבלנו כי המודל הטוב ביותר שהניב לנו את הרווח הגבוה ביותר הוא Random forest , עם 500 עצים וללא המשתנה SP . הרווח הממצוע מנתוני הTest של המודל עומד על 3.433$ ללקוח. משמע אם החברה תחליט ליישם את המודל על 8.2 מיליון הלקוחות שלה, היא צפויה להניב בקירוב רווח של 28,150,000$.

כמו כן, ניתן לשים לב, כי ככל שמגדילים את מספר העצים במודל ה Random Forest מקבלים מודל חיזוי טוב יותר. כיוון ואת הפרויקט אנו עושים במחשב הביתי ניתן להסיק כי שימוש במחשב חזק יותר , אשר יאפשר להריץ את המודל עם מספר עצים גדול יותר , יניב תוצאות טובות אף יותר.

**נספח א'**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **מספר מודל** | **תיאור שלבי העבודה והמודל בקצרה** | **תוצאות המודל על סמך מדד שנבחר** | **הפניה לנתוני ההרצה** | **הערות /מסקנות** |
| 1 | מודל פשוט  ( סיווג של הכל כ"לא ישכיר" ) | 93,004- | A | תוצאה חלשה |
| 2 | הרצת אלגוריתם Random Forest עם 100 עצים, ללא משתנה Reviews | 15,628 | B.1 | תוצאה חלשה |
| 3 | הרצת אלגוריתם Random Forest עם 500 עצים, ללא משתנה Reviews | 19,138 | B.2 | תוצאה חלשה |
| 4 | הרצת אלגוריתם Random Forest עם 500 עצים, ללא משתנה Status Platinum  וללא משתנה Reviews | 18,614 | B.3 | תוצאה חלשה |
| 5 | הרצת אלגוריתם רגרסיה לוגיסטית עם כלל משתני המודל , וללא הוספת משתנה Reviews | 8,550 | C.1 | תוצאה חלשה |
| 6 | הרצת אלגוריתם רגרסיה לוגיסטית ללא המשתנה Status Platinum וללא משתנה Reviews | 9,726 | C.2 | תוצאה חלשה |
| 7 | הרצת אלגוריתם רגרסיה לוגיסטית ללא המשתנה Status Gold וללא משתנה Reviews | 10,096 | C.3 | תוצאה חלשה |
| 8 | הרצת אלגוריתם רגרסיה לוגיסטית ללא המשתנה Status Platinum , ללא המשתנה Status Gold וללא משתנה Reviews | 10,298 | C.4 | תוצאה חלשה |
| 9 | הרצת אלגוריתם רגרסיה לוגיסטית ללא המשתנה Status Platinum וללא המשתנה Status Gold ,ללא המשתנה Call Center Flag וללא משתנה Reviews | 10,140 | C.5 | תוצאה חלשה |
| 10 | הרצת אלגוריתם Naïve Bayes עם כלל המשתנים ללא הוספת משתנה Reviews . | -90,734 | D.1 | Over fitting |
| 11 | הרצת אלגוריתם Naïve Bayes ללא משתנה KM\_L\_Y1 וללא הוספת משתנה Reviews . | -5,270 | D.2 | Over fitting |
| 12 | הרצת אלגוריתם Naïve Bayes ללא משתנה KM\_L\_Y1 , ללא משתנה Status Platinum וללא הוספת משתנה Reviews . | 1092 | D.3 | תוצאה חלשה |
| 13 | הרצת אלגוריתם Naïve Bayes ללא משתנה KM\_L\_Y1 (שנמצא כגורם לבעיית ה Over fitting), ללא משתנה Status Gold ,ללא משתנה Status Platinum וללא הוספת משתנה Reviews . | 1,756 | D.4 | תוצאה חלשה |
| 14 | הרצת אלגוריתם Naïve Bayes ללא משתנה KM\_L\_Y1 ללא משתנה Status Gold ,ללא משתנה Status Platinum , ללא משתנה Call Center Flag וללא הוספת משתנה Reviews . | 6,602 | D.5 | תוצאה חלשה |
| 15 | הרצת אלגוריתם Naïve Bayes ללא משתנה KM\_L\_Y1 ,ללא משתנה Status Gold ,ללא משתנה Status Platinum , ללא משתנה Status Silver ללא משתנה Call Center Flag וללא הוספת משתנה Reviews . | 9,976 | D.6 | תוצאה חלשה |
| 16 | הרצת אלגוריתם Naïve Bayes ללא משתנה KM\_L\_Y1 (שנמצא כגורם לבעיית ה Over fitting), ללא משתנה Status Gold ,ללא משתנה Status Platinum , ללא משתנה Status Silver ללא משתנה Call Center Flag, ללא משתנה VIP Server וללא הוספת משתנה Reviews . | 10,064 | D.7 | תוצאה חלשה |
| 17 | הרצת אלגוריתם Naïve Bayes ללא משתנה KM\_L\_Y1 (שנמצא כגורם לבעיית ה Over fitting), ללא משתנה Status Gold ,ללא משתנה Status Platinum , ללא משתנה Status Silver ללא משתנה Call Center Flag, ללא משתנה VIP Server , ללא משתנה Delay Car וללא הוספת משתנה Reviews . | 10,404 | D.8 | תוצאה חלשה |
| 18 | הרצת אלגוריתם Naïve Bayes ללא משתנה KM\_L\_Y1 (שנמצא כגורם לבעיית ה Over fitting), ללא משתנה Status Gold ,ללא משתנה Status Platinum , ללא משתנה Status Silver ללא משתנה Call Center Flag, ללא משתנה VIP Server , ללא משתנה Delay Car ,וללא משתנה Change Car וללא הוספת משתנה Reviews . | -2,300 | D.9 | תוצאה חלשה |
| 19 | הרצת אלגוריתם Random Forest עם 100 עצים עם משתנה Reviews | 20,456 | E.1 | תוצאה חלשה |
| 20 | הרצת אלגוריתם Random Forest עם 500 עצים עם משתנה Reviews | 22,504 | E.2 | תוצאה חלשה |
| 21 | הרצת אלגוריתם Random Forest עם 500 עצים, ללא משתנה Status Platinum עם משתנה Reviews | **22,536** | E.3 | **המודל הנבחר** |
| 22 | הרצת אלגוריתם Random Forest עם 500 עצים, ללא משתנה Status Platinum וללא משתנה Status Gold עם משתנה Reviews | 22,004 | E.4 | תוצאה חלשה |
| 23 | הרצת אלגוריתם רגרסיה לוגיסטית עם כלל משתני המודל , עם משתנה Reviews | 9,868 | F.1 | תוצאה חלשה |
| 24 | הרצת אלגוריתם רגרסיה לוגיסטית ללא המשתנה Status Platinum ,עם משתנה Reviews | 11,086 | F.2 | תוצאה חלשה |
| 25 | הרצת אלגוריתם רגרסיה לוגיסטית ללא המשתנה Status Gold עם משתנה Reviews | 11434 | F.3 | תוצאה חלשה |
| 26 | הרצת אלגוריתם רגרסיה לוגיסטית ללא המשתנה Status Platinum וללא המשתנה Status Gold עם משתנה Reviews | 11,456 | F.4 | תוצאה חלשה |
| 27 | הרצת אלגוריתם רגרסיה לוגיסטית ללא המשתנה Status Platinum וללא המשתנה Status Gold ,ללא המשתנה Call Center Flag ,עם משתנה Reviews | 11,638 | F.5 | תוצאה חלשה |
| 28 | הרצת אלגוריתם Naïve Bayes ללא משתנה KM\_L\_Y1 עם הוספת משתנה Reviews . | 5,214 | G.1 | תוצאה חלשה |
| 29 | הרצת אלגוריתם Naïve Bayes ללא משתנה KM\_L\_Y1 ללא משתנה Status Platinum עם הוספת משתנה Reviews . | 9,224 | G.2 | תוצאה חלשה |
| 30 | הרצת אלגוריתם Naïve Bayes ללא משתנה KM\_L\_Y1 ללא משתנה Status Gold עם הוספת משתנה Reviews . | 5,636 | G.3 | תוצאה חלשה |
| 31 | הרצת אלגוריתם Naïve Bayes ללא משתנה KM\_L\_Y1 ,ללא משתנה Status Gold ,ללא משתנה Status Platinum עם הוספת משתנה Reviews . | 7,856 | G.4 | תוצאה חלשה |