**סדנה במדעי המידע**

**מטרת הפרויקט**

**חיזוי כמות לייקים של דפי פייסבוק בהתבסס על דפוסי הפעילות בהם.**

פייסבוק הינו ממשק מרכזי כיום לשיתוף ותקשורת חברתיים, הן במגזר הפרטי והן במגזר הציבורי והעסקי. פעילות עמוד פייסבוק נראית פעמים רבות כהולכת בקנה אחד עם פעילות האדם, הגוף והעסק אותו העמוד מתעד. עסקים רבים, גופים ציבוריים, עמותות וכדומה בוחרים להשתמש בממשק זה ורואים בדף הפייסבוק חלון ראווה לפעילותם, השקפותיהם ומטרותיהם. לצד זאת, אנשים פרטיים משקפים דרך עמוד הפייסבוק את מהלך חייהם, חוויותיהם, דעותיהם והשקפותיהם ופרטים אישיים רבים.

בשני המקרים, הפעילות הדף הוירטואלי מהווה מראה לעולם הגשמי, והיא משתקפת בנקודות עיקריות אלה:

פרסומים ודיווחים (Posts) ותגובות חיצוניות לדיווחים אלה, תגובות של בעל הדף לפוסטים חיצוניים, התמונות המפורסמות, המקומות בהם ביקר בעל הדף וכו'. בכל אלה גם משוקפים מאפיינים נוספים, כמו **כמות** הפרסומים, **אורכם**, **אופיים** וכדומה.

בתחילת הדרך, ניסנו לחזות הצלחה וכישלון של עסקים מקומיים על בסיס דף הפייסבוק שלהם. לאור קשיים בתיוגם של עסקים שנכשלו, בחרנו לבסוף בעצת המנחים להתרכז במרכיב עיקרי של עמוד הפייסבוק – כמות הלייקים, שכן זוהי מחווה רווחת ביותר ומרכזית בפייסבוק, והיא משקפת בצורה ברורה את הפעילות של הדף, את התמיכה בו ואת העניין של משתמשי פייסבוק בו. מחקרים רבים נערכו בנושא על מנת לבחון כיצד ניתן להעלות את כמות ה-likes עבור עמוד, מתוך הבנה שעמוד פייסבוק מוצלח ואהוד אשר מגיע להרבה אנשים ביכולתו לקדם לאין שעור את ה"עסק". ננסה לעמוד על המרכיבים העיקריים המשקפים את פעילות והצלחת עמוד הפייסבוק דרך קישורם והשפעתם על **כמות הלייקים** לה זוכה הדף. כמו כן, נבחן את ההשערות והמסקנות עליהם מתבססים המחקרים השונים.

**איסוף המידע**

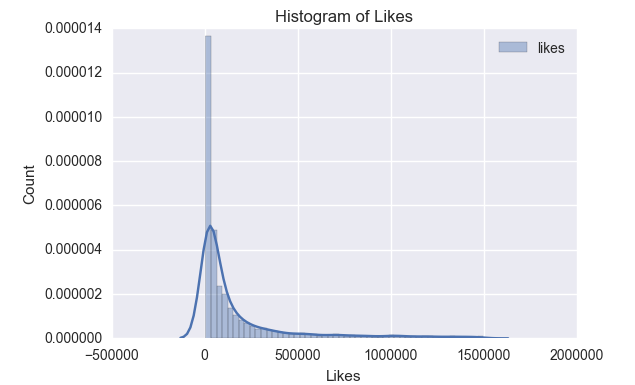
לצורך הבאת המידע על פעילותם של דפי פייסבוק ציבוריים (כגון עסקים, גופים ציבוריים, עמותות, להקות, וכדומה), עשינו שימוש Facebook API ([https://developers.facebook.com](https://developers.facebook.com/)). מאחר וקצב שליפת המידע מוגבל, היה עלינו לאסוף את המידע לאורך זמן רב תוך כדי ניתוחו ועריכתו, תוך עיבוי הולך וגדל של שדות המידע הכלולים בו. בשלב ראשון, ביצענו חיפוש של דפים על בסיס קטגוריות חיפוש, לאחר מכן הבאנו את המידע הכללי על הדף (שם הדף, כתובת, שעות פעילות, פרטי התקשרות וכו'), ואז את הפוסטים ולבסוף את התמונות של הדף. היות וכמות המידע אותה ניתן לשלוף מוגבלת, הבאנו לכל דף את 150 הפוסטים האחרונים ואת 150 התמונות האחרונות. השליפה התבצעה באמצעות facebook client שהרמנו בעוד שהקריאות, שמירת המידע ועיבוד ראשוני שלו נעשו בזאפלין (zeppelin apache, <https://zeppelin.apache.org/> ). בסופו של תהליך ארוך שכלל מהמורות רבות היה בידינו מידע על מעל ל-12,000 דפי פייסבוק שונים.

מסקנה: להגדיל את הכמות.

**תובנות ראשוניות מהנתונים**

המאפיינים השונים לפעילות עמוד פייסבוק, וביניהם כמות הלייקים הגלובלי של הדף, הינם משתנים המתפלגים אקספוננציאלית.

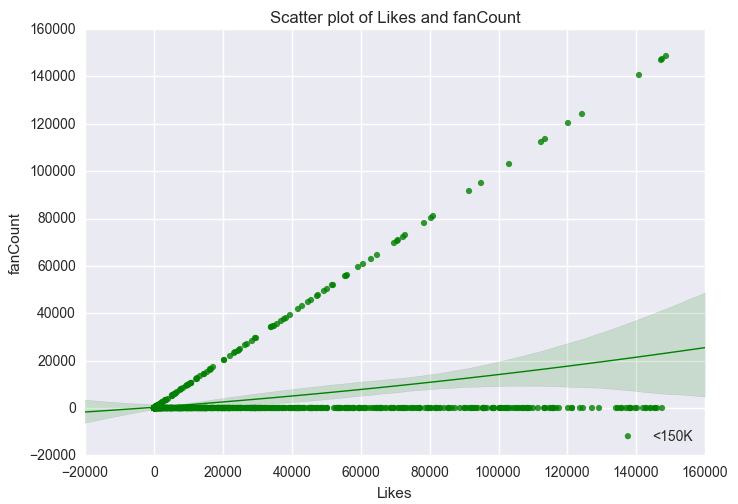
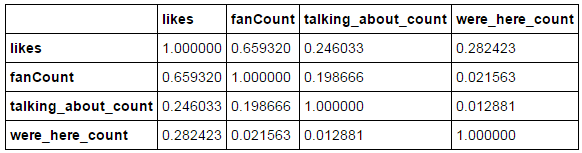
עבור משתנה כמות הלייקים, המקסימום עומד על כ- 10^8 לייקים, והתוחלת על כ- 400,000.



ניתן לשער כי משתנים אשר ילכו יד ביד עם כמות הלייקים הינם מספר המבקרים בעמוד, ומספר העוקבים.

כמו כן, כפועל יוצא מזה, יצטרפו אליהם גם כמות התגובות, התמונות, האירועים וכו'.

ממפת הקורלציות הבסיסית (לפי מקדם pearson) שלהלן ניתן להבחין כי למשתנה ה- fan count קורלציה גבוהה עם כמות הלייקים: לעדכן את הטבלה



**עיבוד מקדים וניקוי המידע**

מאחר והמידע נאסף על ידינו, הרי ששלבים נרחבים בסידור ובסינון המידע נעשה מחוץ למחברת הקוד.

פעולות עיקריות:

* המרת ערכים קטגוריאליים לערכים בוליאניים (dummies variables), דוגמת "מדינה", "כתובת מייל", "אתר" וכדומה. הרוב המוחלט של הדפים מגיע מארה"ב.
* קוונטיזציה של מספר הלייקים ל 4 ערכים דיסקרטיים המתפלגים אחיד. לעדכן את הערכים

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **BIN** | **0** | **1** | **2** | **3** |
| **avg #Likes** | 1166.53 | 12321.80 | 66049.96 | 1594642.00 |

* פיצול המידע ל- Train - מידע המיועד ללימוד המכונה (80%), ו-Train - מידע לבחינת הצלחת החיזוי (20%).
* החלפת ערכים חסרים בערכים חציוניים.
* יצירת משתנים חדשים, דוגמת **צפיפות** **פרסום תמונות**= סה"כ מספר תמונות/תקופת פעילות כוללת.
* הפיכת ערכים טקסטואליים לערכים מספריים (דוגמת שדה "מידע כללי על הדף" ==> "אורך מידע כללי על הדף) ולאחר מכן זריקת ערכים טקסטואליים מחוץ למסד הנתונים.

Data cleaning

**ניקוי על פי סטיית תקן**

מאחר ואנו מתעסקים עם כמויות גדולות של מידע ופיצ'רים רבים אשר הוחזרו לנו על ידי facebook, נרצה ראשית לנקות משתנים אשר ערכיהם קבועים. נעשה זאת על ידי חישוב סטיית תקן, וניקוי features בעלי סטיית תקן של 0.

כבר כאן ניתן לראות כי features אלה תואמים את המצופה, למשל סטיית התקן של שדה has\_profile\_photo הינה 0, שכן לכולם יש תמונת פרופיל.

**ניקוי על פי קורלציה**

משתנים ביניהם יש קורלציה גבוהה מהווים יתירות ב-data set. נרצה לבחון אילו משתנים הינם בעלי קורלציה גבוהה, ולהחליט האם הם מתיישרים עם ההיגיון, האם חלקם מיותרים, והאם אחד עדיף על השני.

למשל- ישנה קורלציה מושלמת בין מספר הפוסטים ש**נכתבו** לבין מספר הפוסטים אשר **פורסמו בפועל**. אין צורך בשניהם.

כמו כן, כמות התמונות באלבומים השונים הינה בעלת קורלציה גבוהה לכמות התמונות הגלובלית.

בחינת ה-data

ננסה לאפיין את המשתנים השונים המרכיבים את ה- dataset שלנו, על פי חקירת ההתפלגויות, הצגתם באופן ויזואלי והצגת מאפיינים של ההתפלגות באופן מספרי.

כאן צריך להוסיף תמונות של ההתפלגויות

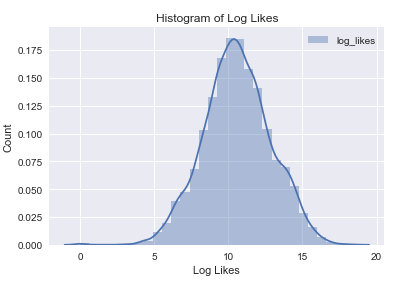
ניתן לראות כי ה- feature אותו אנו מנסים לחזות מתפלג אקספוננציאלית, וכמותו מתפלגים features נוספים חשובים, כמו מספר הפוסטים, מספר התמונות, מספר התגובות וכדומה.

סינון ערכי קיצון

נרצה לנקות ערכים קיצוניים של ההתפלגות, על מנת למנוע הטיות לא רצויות אשר אינן מאפיינות את ה-data.

על מנת להסביר את אופן הביצוע של ניקוי זה, נציג זאת עבור משתנה ה- likes.

המשתנה כאמור, מתפלג אקספוננציאלית, ולכן נבחר להפעיל את פונקצית ה- log על ערכיו. פונקציית ה-log הינה הופכית לפונקציית האקספוננט, ועל כן נצפה לקבל התפלגות אחידה יותר, ולכל הפחות כזו אשר התוחלת שלה הינה גם הערך השכיח שלה:



התקבלה התפלגות בצורת פעמון גאוסי, ואכן הערך השכיח ביותר והערך הממוצע מתאחדים.

כעת, נוכל לבחור לנקות את ערכי הקיצון על ידי ניקוי ערכים הננמצאים מרחק מסויים של סטיות תקן, אותו נגדיר עבור כל אחד מהמשתנים.

נרצה לאפיין את חוסר הסימטריה של הפעמון על מנת שלא לשנות את מאפייניו בעת ניקוי outliers.

נעשה זאת על פי המומנטים מסדר שלישי ורביעי של ההתפלגות- מדדי ה- skewness וה- curtosis.

במקרה זה לא נראית הטיה משמעותית של הפעמון ועל כן ניקוי ערכי הקיצון ישפיעו מעט מאוד על הפרמרטים המרכזיים של ההתפלגות.

**קורלציות**

נרצה לבחון את הקורולציות של המשנים השונים עם ה-feature המרכזי אותו אנו מנסים לחזות.

לכן, נחשב את הקורלציות ונמיינם בסדר יורד, על ידי חישוב המדדים של pearson ושל spearman.

רשימת ערכי הקורלציות

לעומת מדד pearson, אשר מוגבל להתאמה לינארית, מדד spearman מאפיין את ההתאמה באופן שאיננו מוגבל להתאמה לינארית, והרי שאפיון זה נכון יותר בשלב זה.

נבחר את 5 המשתנים בעלי הקורלציה הגבוהה ביותר ונסדרת באופן מטריציוני, על מנת לבחון את הקרוס-קורלציה ביניהם.

מטריצת קרוס קורלציה

אנו רואים כי ישנו מספר משמעותי של features אשר להם קורלציה > 0.5, ועל כן צפוי כי ערך זה יחסום מלמטה את טיב תוצאות ההרצות של המודלים השונים.

ל- features המשמעותיים קרוס-קורלציה של ~0.6. עובדה זו **הינה משמעותית**, שכן הדבר אומר כי הם אינם יוצרים יתירות מיותרת ב-data אחד ביחס לשני, ועל כן שילובם בדרכים שונות יכול להביא לבניית חזאי חזק.

**נרמול**

טווח הערכים של features שונים צפויים להיות פעמים רבות חסרי קורלציה. כמו כן, עבור משתנים רבים הערכים האבסולוטיים אינם אלה המכילים את האינפורמציה המשמעותית, כי אם ערכיהם ביחס למשתנים אחרים. מסיבות אלה, וכן על מנת להתאים את טווח הערכים כך שיתאימו כקלט למודלים השונים, נרצה לנרמל את הערכים.

הנרמול הינו מכמה סוגים:

1. נרמול משתנה ביחס למשתנה אחר – לדוגמה, מספר ה-likes הכולל עבור התמונות אינו מכיל אינפורמציה רבה בפני עצמו שכן הוא איננו בר השוואה בין עמודי פייסבוק שונים. על כן, ננרמלו ביחס למספר התמונות הכולל. היחס הוא זה אשר מהווה המידע המשמעותי, שכן המספר האבסולוטי של ה- likes עבור תמונות תלוי באופן ישיר במספר האבסולוטי של התמונות שפורסמו. נרמול מסוג זה יביא לתוצאות בטווח [0,1].
2. נרמול משתנה ביחס לעצמו – ננרמל את המשתנה ביחס למקסימום שלו, כך שערך המינימום יקבל ערך 0, וערך המקסימום יקבל ערך 1.

מאחר ורוב המשתנים המספריים מתפלגים אקספוננציאלית, נבחן את האפשרות להפעיל עליהם log.

ניתן לראות כי ביצוע פעולה זו מביאה להתפלגות נורמלית!

הצגות כמה היסטוגרמות

ננרמל משתנים גאוסיים אלה (z-normalization) על מנת למרכזם סביב האפס. כך, נכון יותר יהיה להשוות ביניהם ולהכניסם כקלט למודלים הרגרסיה הלינארית וה- LASO. על מנת לשמור על הקונוונציה של מרכוז סביב האפס, נמיר את המשתנים הבינאריים לערכים מספריים של (-1,1) במקום (0,1).

**סידור ה-data לקראת הרצת המודלים**

1. סידור משתנים כ- dummies variables- נפרוש את הקטגוריות של משתנים קטגוריאליים, ונהפכם למשתנים בינאריים, על מנת לתת לקטגוריות משמעות מספרית עבור המודלים.
2. חלוקת משתנה ה- likes לארבעה bins שווי הסתברות לקראת חיזוי מספר ה-bin עבור עמוד הפייסבוק. באופן טבעי, הערכים המאפיינים כל bin (כגון ממוצע, אחוזונים) שומרים על ההתפלגות האקספוננציאלית, שכן החלוקה נעשתה באופן שווה הסתברות.

להעתיק את תוצאות ההרצה של החלוקה ל-bins

**הצגת תוצאות ההרצה כ- confusion matrix**

ביצענו קוונטיזציה של משתנה ה- likes לארבעה bins, וכעת נרצה לבחון האם תוצאות ההרצה הניבו חיזוי נכון.

נציג את התוצאות בצורה מטריציונית אשר תראה את נכונות החיזוי. ציר ה-x (עמודות) מייצג את ה-bin הרצוי, וציר הy (שורות) מייצג את ה-bin אשר נחזה על ידי המודל.

ערכי צבע בהירים מייצגים ערכים גבוהים.

נרצה לקבל מרכוז של ערכי צבע בהירים כמה שיותר סביב האלכסון הראשי של המטריצה, שכן הדבר יצביע על התאמה גבוהה בין ה-bin הרצוי ל-bin אשר נחזה.