דוח הגשה לפרויקט למידת מכונה  
  
**נושא הפרויקט: זיהוי וניתוח דפוסי התנהלות הצרכנים והשפעתם על רכישות והחלטות רכישה**

**מבוא**

בעולם העסקים והשיווק, הבנת התנהגות הצרכנים והשפעתם על החלטות הרכישה היא מרכזית להצלחת כל עסק. בעזרת הבנה עמוקה של דפוסי התנהלות הצרכנים והגורמים המשפיעים עליהם, ניתן לפתח אסטרטגיות שיווק ופיתוח מוצרים שתואמות בצורה מיטבית את צרכי השוק ומשפרות את תהליכי המכירה.

פרויקט זה מתמקד בשימוש בטכניקות מתקדמות של למידת מכונה כדי לזהות דפוסים בהתנהלות הצרכנית, כך שניתן יהיה להבין ולנתח מהם המשתנים המרכזיים המשפיעים על התנהלות הצרכנים בתהליך הרכישה והשימוש במוצרים. על ידי זיהוי דפוסים והבנת הגורמים המשפיעים על התנהלות הצרכנים, ניתן יהיה לפענח ולהבין אילו גורמים מניבים רווחיות גבוהה ואילו עשויים להפחית ממנה.

מטרת הפרויקט היא לספק כלי ניתוח והבנה של התנהלות הצרכנים, המאפשרים לחברות לקבל החלטות מבוססות נתונים ומידע מדויק, ולפתח אסטרטגיות שיווק ופיתוח מוצרים בהתאם לצרכי הלקוחות. באמצעות הבנת הגורמים המשפיעים על התנהלות הצרכנים, ניתן יהיה לייעל את תהליכי השיווק.

**דאטה סט ופיצ'רים**

בדאטה שלי השתמשתי בקריאה לAPI מאתר https://catalog.data.gov/dataset/  
העוסק בצריכת לקוחות.  
העמודות הרלוונטיות שלקחתי מהדאטה לאחר ניקיון לפי רלוונטיות הן:

הסבר על כל עמודה בטבלה

sid: מזהה ייחודי לכל שורה בטבלה.

id: מזהה ייחודי לכל שורה בטבלה, יכול לשמש כמפתח ראשי.

State FIPS code: מזהה ייחודי של מדינת ארה"ב, המשמש במערכת הפדרלית של ארצות הברית לצורך זיהוי גאוגרפי של המדינות.

Date: תאריך השורה.

All merchant category codes spending: הוצאות בקטגוריות השונות של רשתות המסחר.

Accommodation and food service (ACF) spending: הוצאות בתחום האירוח והשירותים המזון.

Arts, entertainment, and recreation (AER) spending: הוצאות בתחום האמנות, הבידור והנפנוף.

General merchandise stores (GEN) and apparel and accessories (AAP) spending: הוצאות בחנויות למוצרים כלליים, בגדים ואביזרים.

Grocery and food store (GRF) spending: הוצאות בחנויות מזון ומכולת.

Health care and social assistance (HCS) spending: הוצאות בתחום הבריאות והסיוע החברתי.

Transportation and warehousing (TWS) spending: הוצאות בתחום התחבורה והאחסון.

Retail spending, including grocery (AAP, CEC, GEN, GRF, HIC, ETC, SGH): סכום ההוצאות בתחום המסחר, כולל קניות בחנויות מזון.

Retail spending, excluding grocery ((AAP, CEC, GEN, HIC, ETC, SGH): סכום ההוצאות בתחום המסחר, ללא קניות בחנויות מזון

העמודות האלה נבחרו לאחר ניקוי.  
הסיבה שבחרתי להסיר חלק מהעמודות לחלק הראשון של הפרויקט היה משום שהן היו מיותרות ואסביר מדוע:  
את העמודה meta הורדתי כי התוצאה של table['meta'].unique() היתה מערך ריק.  
את העמודה position הורדתי משום שהערך בה שווה ל0 בלבד ואינו משפיע כרגע על המשך הפרויקט.  
את העמודות 'created\_at', 'updated\_at' הורדתי גם כן משום שערכן הבודד הזהה לאורך הטבלה אינו מוסיף להמשך הפרויקט.  
created\_meta, updated\_meta – גם כן הורדתי משום שלא ראיתי שום תועלת בתוכן שלהן.  
כלומר מ19 עמודות נותרו 13 אשר היו רלוונטיות למטרת הפרויקט.

את השוני והתועלת בין עמודות id לsid בדקתי ולאחר מחשבה החלטתי להשאיר למקרה שאולי יועילו משום שאינן מיותרות לחלוטין ויש עניין להשאירן.  
  
בהמשך רציתי לבדוק התפלגות של קודי מדינה על מנת להסיק אזורים שכיחים לפרויקט:  
תמונה שמכילה מלבן, תרשים, קו, צילום מסך

התיאור נוצר באופן אוטומטי  
  
בהמשך הפעלתי describe על מספר תכונות על מנת לראות איזה תכונה אשר אני מניחה שיכולה להשפיע אכן משפיעה מבחינת הנתונים.  
העמודה "All merchant category codes spending" הייתה ניכרת במדדים שלה.  
  
לאחר מכן בחנתי בגרף את התפלגות ההוצאות הכוללת קניות והעברתי קו המייצג את הממוצע.  
תמונה שמכילה פיקסל

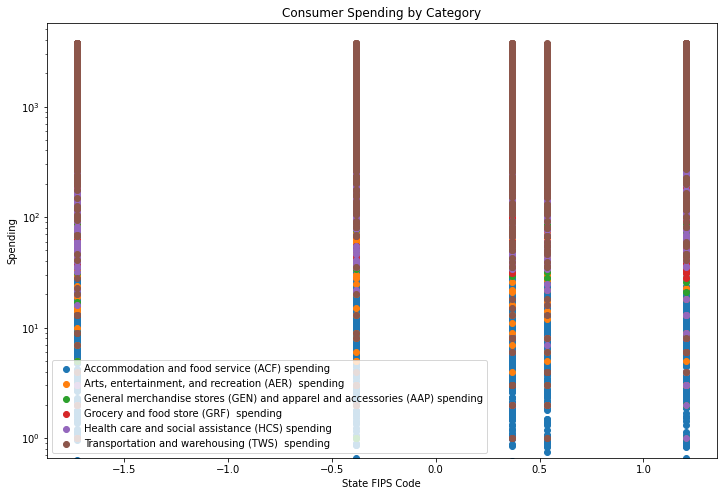
התיאור נוצר באופן אוטומטי  
הקוד מייצר היסטוגרמה של נתונים מעמודה ההוצאות הכוללת קניות בטבלה, מבצע ניקוי וסידור ערכים, חישוב ממוצע, ומציג את ההיסטוגרמה עם קו שמייצג את הממוצע.

בגרפים שלאחר מכן בדקתי את התפלגות הנתונים להוצאות עם או ללא קניות על מנת להגיע לדפוס מסויים של דאטה אותו אוכל לחקור.

לנתונים ששלפתי הדאטה הבנתי שהמחקר המתאים לבצע עליהם יהיה באמצעות רגרסיה.

חילקתי את הנתונים לקבוצת אימון ובדיקה וחישבתי את השגיאה mse.. השימוש בפונקציה מאפשרת להפריד את נתוני הרגרסיה לקבוצת אימון וקבוצת בדיקה.

לאחר מכן ביצעתי אימון של רשת נוירונים עבור בעיה של חיזוי. באמצעות האימון הזה ניתן להעריך את ביצועי המודל ע"י התאמת הפרמטרים באופן מדויק.

לאחר מכן המשכתי לבדוק על מה צרכנים מוציאים את כספם לפי קטגוריה:  


הגרף הוא התפלגות הוצאות הצרכנים לפי קטגוריות שונות ביחס לקוד המדינה -

כל נקודה בגרף מייצגת ערך של קטגורית הוצאה מסוימת במדינה מסוימת.

על פי הגרף, ניתן לראות את ההבדלים בהוצאות הצרכנים בין המדינות.

הגרף משתמש בקנה מידה מסויים כדי להציג טווח רחב יותר של הערכים בציר ה-Y (הוצאות), וזאת משום שהערכים בקטגוריות השונות עשויים להיות בטווחים שונים ורחבים. כל נקודה בגרף מייצגת את רמת הוצאות הממוצעת של קטגוריה מסוימת בכל מדינה. השימוש בקנה מידה מאפשר להבין את התפלגות ההוצאות בצורה יעילה, גם כאשר ישנם ערכים גבוהים או נמוכים יחסית.

לאחר מכן המשכתי התאמה במודלים שונים. כאשר בדקתי דיוק בעמודות

X = data[['State FIPS code']]

y = data['Health care and social assistance (HCS) spending']

לפי הדיוק (accuracy) של המודל שנוצר הוא כ-0.2%, מה שאומר שהמודל לא מצליח לזהות באופן יעיל את הקשר בין 'State FIPS Code' לעמודה של 'Health care and .social assistance (HCS) spending'. תוצאה נמוכה מאוד היכולה להצביע על כך שהמודל לא יעיל בזיהוי הקשר בין המשתנים בנתונים האלה.

.

לעומת זאת כשבדקתי את בחירת העמודות הרלוונטיות עמודת הוצאות בריאות ורווחה ועמודת קוד המדינה

X = updated\_table[['Health care and social assistance (HCS) spending']]

y = updated\_table['State FIPS code']

ערך ה-MSE שקיבלנו הוא כ-0.96. ערך זה מייצג את המרחק הרבועי הממוצע שבין הערכים הניבויים על ידי המודל לבין הערכים האמיתיים בקבוצת הבדיקה.

ערכים נמוכים של MSE מעידים לנו על טווח טוב יותר של החיזוי ככל שהוא נמוך יותר, כך החיזויים יתקרבו יותר לערכים האמיתיים.

במקרה זה, ערך ה-MSE הוא נמוך, מה שמעיד על טווח צפוי טוב (יחסית) של החיזויים של המודל. מכך אפשר להסיק שהמודל עשוי להיות יעיל בחיזוי ערכי 'State FIPS code' בהתבסס על הוצאות בריאות ורווחה.

על מנת לשפר את הפרויקט יש הרבה ניסוי וטעיה עד שנגיע אל הנתונים הרלוונטים.  
  
לאחר בדיקה של מודלים מצאתי שהערכים

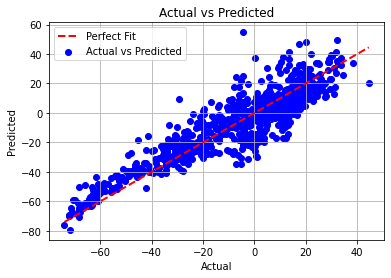
X = updated\_table[['All merchant category codes spending', 'Retail spending including grocery', 'Retail spending excluding grocery']]

y = updated\_table['Accommodation and food service (ACF) spending']  
  
רלוונטיים לי להמשך הפרויקט ויכולים לקדם את המטרה בהבנת ההוצאות.  
בבדיקה בהתאם של מודל רגרסיה לינארית יצאו לי הערכים:

Mean Squared Error: 108.67090179952714

Root Mean Squared Error: 10.424533649018892

R^2 Score: 0.8184114549465873



מספר מסקנות:

1. Mean Squared Error (MSE - שגיאה ריבועית ממוצעת): ערך זה מציין את השגיאה הממוצעת שבין הערכים האמיתיים לבין הערכים שנחזיק על ידי המודל. במקרה שלך, ה-MSE הוא 108.48. ערך זה מציין שהשגיאה הממוצעת בין הערכים האמיתיים לבין החיזויים היא כ-10.415.

2. Root Mean Squared Error (RMSE - שגיאה ריבועית ממוצעת בריבוע): זהו השורש של ה-MSE ומשמש כמדד של השגיאה הממוצעת בין הערכים האמיתיים לבין החיזויים. במקרה זה, ה-RMSE הוא 10.415, מה שאומר שהשגיאה הממוצעת בין הערכים האמיתיים לבין החיזויים היא כ-10.415.

3. R^2 Score (: זהו המדד שמציין כמה הנתונים מתארים את הדפוסים של הנתונים האמיתיים. בעזרתו ניתן להעריך כמה המודל הזה מתאר את הנתונים. במקרה שלי, ערך R^2 הוא 0.8209, שהוא גבוה מאוד. זה אומר שהמודל מצליח להסביר כ-82.09% מהשונות בנתונים, כלומר טוב לנו.

הגרף מציג את הנקודות האמיתיות מהנתונים הבדיקה על הציר האופקי (x-axis) והערכים שנחזו על ידי המודל על הציר האנכי (y-axis). כל נקודה בגרף מייצגת זוג ערכים - הערך האמיתי מהנתונים הבדיקה והערך שנחזה על ידי המודל.

הקו האדום המקווקו מייצג קו ההתאמה המושלם, כאשר עבור כל נקודה אמיתית על הציר האופקי, יש נקודה על הקו האדום עם אותה ערך על הציר האנכי.

אם הנקודות נמצאות קרובות לקו האדום, זה מראה על דיוק גבוה של המודל. במקרה הזה, יש לנו התאמה טובה בין הערכים האמיתיים לבין הערכים שנחזו על ידי המודל.

בשלב השני, אפעיל אלגוריתמים של clustering כדי לקבץ את הצרכנים לקבוצות דומות

בהתבסס על דפוסי התנהלותם. אשתמש בטכניקות כמו means-K או DBSCAN על מנת

לקבוץ ולזהות את הקבוצות השונות ביותר.ראש הטופס

הקוד בוחר את מספר הקבוצות הרצוי לקלסטרינג, במקרה זה נבחר מספר קבוצות שווה ל-5. לאחר מכן, המודל KMeans מתאמן על הנתונים ומבצע את הקלסטרינג. בשלב הבא, נוספה עמודה חדשה לדאטה updated\_table המכילה את מספר הקבוצה לכל רשומה, כך שניתן לראות לאיזו קבוצה הוקצו הצרכנים.  
  


ברגע שלוקחים בחשבון התנהגות צרכנים, (כמטרת הפרויקט), יש לנו צורך להבין איך הם מוציאים את הכסף שלהם (הוצאות בקטגוריות שונות) ובאילו סוגי מקומות הם מבצעים רכישות (קטגוריית אירוח ומזון).

עמודות כמו 'All merchant category codes spending' מציינות את כמות הכסף שהצרכן מוציא בקטגוריה מסויימת, מה שיכול להעיד על העדפות הרכישה שלו. קטגוריית האירוח והמזון חשובה מאוד מבחינת התרשמות הצרכן ממקומות האירוח והאוכל, השירות, המחירים ועוד. לכן, ארצה לראות כיצד הצרכנים מפזרים את ההוצאות שלהם בין הקטגוריות השונות.

בכך, בחירת העמודות 'All merchant category codes spending' ו-'Accommodation and food service (ACF) spending' מתאימה בצורה טבעית למטרה כדי להבין את התנהלותם של הצרכנים ואת השפעתם על החלטות הרכישה שלהם.

לאחר מכן ביצעתי קלאסטריגג באמצעות DBSCAN.

המסקנה מהפלט היא שהנתונים נחלקים לקבוצות דומות בהתבסס על דפוסי התנהלותם של הצרכנים. הפלט מציג את מספר הצרכנים בכל קבוצה ואת המזהים של הקבוצות. לדוגמה, בפלט שצוין ישנן קבוצות שונות, כאשר בקבוצה הראשונה יש 1780 צרכנים, בקבוצה השנייה יש 857 צרכנים, וכן הלאה. המסקנה היא כי הנתונים מצביעים על התפלגות של הצרכנים לקבוצות שונות, כאשר כל קבוצה מייצגת דפוס שונה של התנהלות.המספרים השליליים (-1) מייצגים את הנקודות החריגות או את הנקודות שלא הצליחו להתקבץ לאף קבוצה.זה יכול להיות בגלל שהן מרוחקות מהקבוצות הראשיות ואין להן דמות נכונה בין הנקודות האחרות, או שהן יכולות להיות נקודות אשר לא עברו בהצלחה את תהליך הקלסטרינג.

לאחר מכן ביצעתי מספר מודלים על מנת לחזות את המודל הטוב ביותר לערכים:

# בחירת תכונות המשמשות כמשתנים לחיזוי

X = data[["Retail spending excluding grocery", "Retail spending including grocery", "Accommodation and food service (ACF) spending"]]

# בחירת המשתנה לחיזוי

y = data["All merchant category codes spending"]

השתמשתי ב4 מודלים: מודל רגרסיה לינארית, מודל רשת נוירונים מלאה, מודל רגרסיה פולינומיאלית ו עץ החלטה.  
לפי תוצאות:

Metric Result 1 Result 2 Result 3 Result 4

0 Mean Squared Error 25.419871 22.880303 23.941122 38.838658

1 Mean Absolute Error 3.681657 3.482348 3.567956 4.316484

2 R-squared 0.892217 0.902985 0.898487 0.835320

3 Root Mean Squared Error 5.041812 4.783336 4.892967 6.232067

מהתוצאות אפשר לראות: מתוך התוצאות של המדדים השונים:

1.Mean Squared Error (MSE): הערכים הנמוכים ביותר הם הטובים ביותר. לכן, המודל הטוב ביותר הוא המודל שמקבל ערך נמוך ביותר ל-MSE. כאן, מודל 2 הוא המינימלי עם 22.894153. 2.Mean Absolute Error (MAE): גם כאן, הערכים הנמוכים ביותר הם הטובים ביותר. מודל 2 הוא המינימלי עם 3.345378. 3.R-squared: הערכים הגבוהים ביותר הם הטובים ביותר, כאשר 1 מציין התאמה מושלמת של המודל לנתונים.מודל 2 הוא המירבי עם 0.902926. 4.Root Mean Squared Error (RMSE): כמו MSE, הערכים הנמוכים ביותר הם הטובים ביותר. מודל 2 הוא המינימלי עם 4.784783.

לפיכך, מבחינת כל המדדים, מודל 2 הוא המודל הטוב ביותר לחיזוי.

​