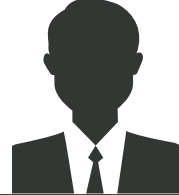




חיזוי ביצועי ציוותי עובדים במשימות טוריות בשווקי העבודה המקוונים



תומר נבע



אביחי שריקי

אביחי שריקי הוא דוקטורנט בפקולטה לניהול ע"ש קולר באוניברסיטת תל אביב, בהנחייתה של פרופ' ענבל יהב. בעל תואר ראשון בכלכלה ולימודי מזרח תיכון מאוניברסיטת חיפה ותואר שני בניהול מערכות מידע מאוניברסיטת תל אביב. תחום המחקר שלו הוא זיהוי משמעותיות בטקסט, במיקוד לשפה העברית. הוא פיתח את HeBERT, מודל שפה עברי המבוסס על ברט ומודל לזיהוי רגשות בעברית מטקסט.

ד"ר תומר נבע הוא חבר סגל בכיר בפקולטה לניהול ע"ש קולר באוניברסיטת תל אביב. בעבר שימש כראש התמחות מדעי הנתונים בניהול ולפני שהצטרף לפקולטה לניהול היה חוקר אורח ב-New York University ופוסט דוקטורנט בגוגל. מחקריו עוסקים בפיתוח אלגוריתמים של למידת מכונה ושילובם עם מידע המתקבל מבני אדם לצורך חיזוי התנהגות, הערכת ביצועי עובדים והערכת יכולות מודל שפה גדולים. מחקריו פורסמו בכתבי עת מובילים בתחומי למידת מכונה, מדעי הנתונים ומערכות מידע. מחקריו גם נתמכו על ידי גופים שונים, ובהם הקרן הלאומית למדע. בעל תואר ראשון בהנדסת תעשייה מהטכניון, ותואר שני ושלישי מהפקולטה לניהול ע"ש קולר באוניברסיטת תל אביב.

תקציר

בעבודה זו התמקדנו בבעיות המתגרות של חיזוי תוצאות שיבוץ עובדים בשוקי העבודה המקוונים. שווקים אלו מאפשרים גיוס רבבות עובדים לתקופת זמן קצרה למשימות מוגדרות. התמקדנו בחיזוי ביצועי צמדי עובדים הפועלים על משימה טורית משותפת בשוקי עבודה מקוונים, כמקרה בסיסי של עבודת צוות בשווקים אלו. באופן ספציפי בחנו את יכולת החיזוי (Predictability) של איכות ביצועי עובדים משותפים המבוססת על עבודת כל עובד בנפרד וגם על הסינרגיה בין העובדים. השערת המחקר היא כי ישנה תרומה ליכולת החיזוי של ביצועי העובדים המשותפים כאשר משתמשים במידע על מאפייני שני העובדים. כלומר, מודל חיזוי המתבסס על וקטור המאפיינים של שני העובדים יאפשר חיזוי טוב יותר ממודל חיזוי המתבססים על כל אחד מהעובדים בנפרד. על מנת לבחון זאת מימשנו ניסוי בפלטפורמת העבודה המקוונת Amazon Mechanical Turk שבו ציוותנו אלפי עובדים לטובת ניתוח משותף של כתבות פיננסיות. על ידי שימוש הציוותים הללו כמקורות מידע, ניסינו לחזות את ביצועי העובדים על בסיס מאפייניהם האישיים, כפי שהפלטפורמה מאפשרת לסנן למעסיקה, ובאמצעות שימוש במידע היסטורי על ביצועי העובדים בחלק מהמקרים. הצלחנו להראות כי בהינתן מידע מועיל בנתונים (כמו נתונים היסטוריים על ביצועי העובד) ושימוש באלגוריתם רגרסיה חזק דיו, קיים שיפור מסוים עבור מודל חיזוי ביצועי הצלחת העובדים המתבסס על נתוני שני העובדים לעומת מודל בסיסי (המסתמך על נתוני אחד העובדים) או נאיבי (המתבסס על ממוצע הצלחת כלל העובדים). עם זאת, השערתנו לא נתמכה במקרה שבו אין נתונים היסטוריים לגבי כל עובד (בעיית Cold start). תרומת העבודה היא בעצם ניסוח בעיית ציוותי העובדים כבעיה עסקית מבוססת נתונים, ובבחינת יכולת החיזוי של ביצועי עובדים משותפים.



(C) Sharon Toker

מבוא

מ-52 מיליון עובדים בשווקים אלו היו רשומים ברחבי העולם (Codagnone et al., 2016), והיקף העבודה המקוונת הוערך בכ-15 עד 25 מיליארד דולר (Kuek et al., 2015).

בעבודה זו ביקשנו לבחון יכולת חיזוי (Predictability) של ביצועי צוותי עובדים המבצעים משימה טורית דו-שלבית. במקרה זה, משימה טורית דו-שלבית משמשת כ"מקרה פרטי" למשימות טוריות ארוכות אף יותר. לדוגמה – ניתוח נתונים פיננסיים (שלב א' של המשימה הטורית) וסיכומים לכדי דוח מנהלים (שלב ב'). לשם כך ברצוננו לבחון האם ניתן לחזות את ביצועי העובדים על בסיס מאפייני שני העובדים באופן טוב יותר ממודל הנסמך על מידע לגבי עובד אחד בלבד. תרומת עבודה זו היא ניסוח בעיית ביצועי הצוות בשווקי העבודה המקוונים כבעיה עסקית מכוונת נתונים (Data driven)

שווקי עבודה מקוונים מייצרים קשרי עובד-מעסיק ייחודיים ואינטראקציה שונה בין אנשים לטכנולוגיה. כך, נוצרו בפועל שוקי עבודה לציבור הרחב (Crowd sourcing) המאפשרים מיקור חוץ של רבבות "עבודות קטנות" (Micro-tasks). למעשה, שוקי העבודה המקוונים מספקים סביבה נוחה לטכנולוגיות מדעי נתונים (Data science) ולמידת מכונה (Machine learning), לניהול עובדים אנושיים ולהגברת היעילות (Ipeirotis, 2010). שיפור היעילות בשווקים אלו קריטי לאור צמיחתם העקבית. לדוגמה, היקף כוח העבודה בפלטפורמת העבודה המקוונת של אמזון לבדה (Amazon Mechanical Turk) מוערך בכ-100 עד 200 אלף עובדים קבועים (Djellel et al., 2018). בסך הכול דווח שיותר

1.2 עבודת צוות בשוקי העבודה המקוונים

עבודה משותפת של עובדים ושיתוף פעולה מקוון הוכרו כאלמנטים חשובים ביותר לשיפור יעילות שוק העבודה המקוון ולתמיכה ביעדים ארגוניים. למשל, Kittur et al. (2013) ציינו כי שיפור שיתוף הפעולה והסנכרון הוא גורם חשוב שיש להתייחס אליו כדי לאפשר לשווקים מקוונים לבצע ביצועים יעילים. Little et al. (2009) הראו כי שיתוף פעולה של עובדי crowd מוביל לתוצאות מהימנות ומהירות יותר מאשר עובד יחיד.

מעבר לכך, חשיבות בניית צוות טוב עולה לאין שיעור בצוותים וירטואליים. למשל, Cramton (2001) ו-Driskell (2003) הראו מקרים שבהם צוותים וירטואליים שעובדים ביחד (במקביל או באופן טורי) עובדים באופן פחות יעיל, מתקשים לייצר למידה משותפת, ובעלי פוטנציאל גבוה יותר ל"הפלת" משימות. מחקרים כגון Li et al. (2011) התמקדו במציאת קבוצת העובדים היעילה ביותר למשימה על בסיס מאפייני העובדים, כמו רמת השכלה, תיכון וגיל. לעומתם, המחקר של Demartini (2013) ניתח אילו עובדים צריכים לבצע משימה נתונה על סמך פרופיל העובדים המופק מרשתות חברתיות כדי להניע ליעילות מרבית (מצד העובד ומצד המעסיק).

המחקר שלנו מתמקד במשימות טוריות המבוצעות על ידי צוותים של עובדים מקוונים. Fidler (2015) הראה שמשימות אלו מספקות יכולות כה חזקות עד שניתן יהיה להחליף עובדים מומחים בכמה עובדים מקוונים בעלי מיומנות נמוכה יותר שיבצעו משימות טוריות. Kittur et al. (2011) הציעו מבנה שבו ניתן להשתמש בפלטפורמות העובדים המקוונים לשם פתרון בעיה מסובכת. כך, הם פירקו את הבעיה לתתי-בעיות ומשימות לפי שלבים, ואת ממצאי כל אחד מהשלבים "העבירו" באופן טורי בין העובדים בשלבים השונים לפתרון מלא. Fidler (2015) הראה כיצד באמצעות שימוש בקבוצת עובדי Crowd ניתן לכתוב דוח שנתי לאחת מחברות Fortune 50 בעשירית מהזמן הממוצע ובאיכות שאינה נופלת מהדוחות הקודמים שלה. כך אפשר לבצע מספר רב של משימות ביעילות על ידי עובדים מקוונים זולים ולחסוך בהוצאות הפרויקט הכולל.

(business problem) ובחינת יכולת החיזוי של ביצועי צוותי עובדים במסגרת ניסוח הבעיה.

מחקרים קודמים הראו כי עבודת צוות יעילה יותר מעבודה יחידנית (Campion et al., 1993) והדגישו את התרומה החשובה של צוותי עבודה מקוונים/וירטואליים (Little et al., 2013; Kittur et al., 2009; al., להשגת ביצועים יעילים. לכן היכולת לחזות את הצלחתם של צוותי עובדים, במקביל לגידול בשוקי העבודה המקוונים, יכולה להעלות את יעילות הצוות (מפני שניתן יהיה לשבץ עובדים מראש בצוותים שהכי מתאימים להם). כמו כן, חיזוי מדויק של הצלחת הצוות יכול לצמצם את העלות למעסיק (הנובעת מתיקון עבודת הצוות, שליחה לצוות אחר וכדומה), וכפועל יוצא להעלות את הערך לחברות ועסקים. יתרה מזאת, פיתוח מתודולוגיה לחיזוי הצלחת הצוות ייתן ערך מוסף לפלטפורמות שוקי העבודה המקוונות – הן כלפי בעלי עסקים הפועלים דרכן והן כלפי העובדים עצמם (על ידי מתן עבודות וצוותים "מותאמים אישית").

1. סקירת ספרות

1.1 עבודה משותפת בארגונים

עבודת צוות היא התהליך שבו חברי הצוות משתפים פעולה כדי להשיג יעדי משימה. עבודת צוות מתייחסת לפעילויות שבאמצעותן תשומות הצוות מתורגמות לתפוקות כמו יעילות ושביעות רצון של הצוות. מחקרים במדעי החברה ובפסיכולוגיה ניסו למדוד את הצלחת הקבוצה והמאפיינים המשפיעים עליה ביותר. כך, הודגם כי עבודת צוות יעילה יותר מעבודה של שני בודדים (Campion et al., 1993, Brannick et al., 2003, Hoegl et al., 2003, Driskell et al., 1995, al.). אולם קיימים גורמים משפיעים על ביצועי הצוות בארגון, ובהם הרכב תכונות הצוות וגיוונו (Susan, et al., 2013), הכשרת הצוות ושיתופי פעולה קודמים (Dow et al., 2017), וסביבת העבודה שבה הצוות עובד (Arsalan et al., 2018). עם זאת, מחקרים אלו ודומיהם התבססו ככלל על מחקר איכותני מסביר המבוסס על משאלים וסקרים. לעומת זאת, בעבודה זו נרצה לחזות את ביצועי צוותי עובדים על בסיס נתונים ומודלי למידת מכונה לחיזוי.

1.3 מחקרים אלגוריתמיים לחיזוי ביצועי עובדים ולציוות עובדים

העבודה שלנו שונה מעבודות אלגוריתמיות קודמות העוסקות בציוות עובדים שלא לקחו בחשבון את הסינרגיה בין העובדים או לקחו אותה כ"נתון פתיחה" ידוע. הן גם הניחו שצוות "טוב" הוא קבוצה (או זוג, כמקרה פרטי) שבה סכום יכולות חברי הצוות מכסה את כל תחומי המשימה (למשל, Golshan, 2014) מבלי לקחת בחשבון את הסינרגיה או האינטראקציה בין העובדים. לחלופין, מחקרים ניסו לבחון כיצד לבצע השמה יעילה של הקבוצה הערכית ביותר תוך מזורז העלויות, מבלי להתחשב כלל באינטראקציה ביו חברי הקבוצה (Aris 2010; Arias et al., 2011; Brocco et al., 2016)

מחקרים אלגוריתמיים אחרים שקלו סינרגיה והפחתת עלויות התיאום אך הניחו רק הגדרה שבה איכשהו ידוע אופן ביצוע משותף. לדוגמה, Dorn and Kargar et al., 2013; Dustdar, 2010, בחנו עובדים שעובדים במשותף באמצעות רשת עבודה. Datta et al. (2011 וב-2012) בנו אלגוריתם להרכבת צוות לפתרון משימה הדרושה מספר כישורים על ידי השמה יעילה של העובדים לפי הכישורים שלהם, ובחינת טיב הקשר בין האנשים על בסיס נתונים מרשתות חברתיות. Colomo-Palacios (2012) ניסו לענות על בעיה דומה אך הניחו כי המידע על העובדים מוזן על ידי המנהלים שלהם. כלומר, כל המחקרים עד כה הניחו כי קיים מידע כלשהו על העובדים, כולל מידע על כישוריהם ומשמעות האינטראקציות שלהם עם עובדים אחרים לגבי הצלחת ביצוע משותף, מידע שסביר להניח כי אינו זמין עבור מעסיקים של עובדי crowd. במחקר שלנו ננסה לענות גם על הפער הזה – נייצר סט מאפיינים מנתונים ציבוריים על העובדים ונבחן את יכולת החיזוי שלהם להצלחת הצוות.

הבעיה שאנו עוסקים בה שונה ממחקרים קודמים גם בשימוש במדעי נתונים לצורך הערכה וניבוי של ביצועי עובדים בודדים, כמו Raykar et al. 2010; Ipeirotis et al. 2014; Geva et al. 2016; Ipeirotis, I and Saar Tsechansky, 2016; Kokkodis et al., 2019; Wang et al., 2017; Geva et al., 2016). מחקרים אלה נועדו להסיק באופן אלגוריתמי או להעריך רטרואספקטיבית את ביצועי העובדים האישיים תחת הגדרות שונות.

בנוסף, ודאי שעבודתנו שונה ממחקרים שניסו להעריך את הצלחת העובדים על בסיס הנחת עובדים כרובוטים

עם מאפיינים ייחודיים הקשורים לתיאום משימות (למשל, Veloso, 2014; Liemhetcharat ו-2014) ולא לבני אדם. במחקרנו ננסה להעריך את ביצועי העובדים כפרט וכצמד וננסה לחזות את ביצועיהם המשותפים. ננסה להראות כי אנו יכולים לעשות זאת גם בלי היסטוריית עבודתם (Cold start).

2. תיאור תהליך בדיקת יכולת החיזוי

2.1 רקע כללי

כאמור, במחקר זה אנו מבקשים לבחון האם ביכולתנו לחזות את ביצועי הצוות בצורה מדויקת יותר על סמך נתוני שני חברי הצוות, ובפרט נרצה לבחון האם חיזוי המתבסס על וקטור המאפיינים של שני העובדים טוב יותר מחיזוי הצלחת הצוות המתבסס על מאפייני כל אחד מהעובדים בנפרד. עוד נרצה לבחון האם ומהי מידת שיפור החיזוי המתקבל משימוש במידע היסטורי על ביצועי העובדים.

באופן ספציפי ברצוננו לבחון יכולת חיזוי של ביצועי צוותי עובדים המבצעים משימה טורית דו-שלבית. במקרה זה משימה טורית דו-שלבית משמשת כ"מקרה פרטי" למשימות טוריות ארוכות אף יותר. לצורך כך ברצוננו לבחון האם ניתן לחזות את ביצועי העובדים על בסיס מאפייני שני העובדים באופן טוב יותר ממודל הנסמך על מידע על עובד אחד בלבד. השערתנו היא כי טעות חיזוי הצלחת הצוות קטנה יותר עבור מודל המקבל נתונים על שני העובדים מאשר על אחד העובדים בלבד.

לצורך כך נבחרה משימת ניתוח סנטימנט (Sentiment) של כתבה חדשותית פיננסית כמשימה טורית דו-שלבית, T, לטובת הניסוי במחקר זה. השלב הראשון במשימה הוגדר כסיכום הכתבה (ממוצע של כ-2,500 מילים) לכדי משפט אחד או שניים והוא בוצע על ידי עובד מסוים. בשלב השני נדרש עובד אחר לדרג את סנטימנט הכתבה – "עד כמה המידע בכתבה הוא חיובי או שלילי לצורכי השקעה", להלן ("ינכונות להשקעה") לטובת משימת השקעה פיננסית על בסיס הסיכום בלבד. תוצר המשימה הוגדר כציון שניתן כתוצאה מהעבודה המשותפת של עובד מסוים משלב אי עם עובד מסוים משלב ב' על כתבה ספציפית.

2.2 איסוף הנתונים לניסוי

השניאה בין ערך "הנכונות להשקעה" האמיתי לבין זה שניתן על ידי עובד שלב ב' בצוות.

מבחינה מעשית, בשלב הראשון של כל משימה בניסוי ביקשנו מכל אחד מ-500 העובדים (מתוך ה-1,000) לסכם 30 כתבות חדשותיות כלכליות שנבחרו באקראי. בסוף השלב הזה היו ברשותנו 15,000 סיכומים של 160 כתבות חדשותיות שונות. בשלב השני ביצענו השמה רנדומלית של 500 עובדים אחרים ("עובדי שלב ב'") ל-15,000 הסיכומים שעובדי שלב א' ביצעו, על מנת שייתנו ציון סנטימנט לסיכומי הכתבות לטובת משימת השקעה פיננסית. בסוף שלב זה היו ברשותנו 15,000 ציוותים שונים בין עובדי שלב א' לשלב ב' עם הערכות הסנטימנט שלהם.

2.3 נתוני אימון ומבחן

חשוב להדגיש שעל מנת שיהיו בידינו נתוני אימון (Training set) ותקפים, נפרדים מנתוני המבחן (Holdout set), פיצלנו באופן מלא את הציוותים לשניים. כך 50% מעובדי השלב הראשון יכלו לעבוד פוטנציאלית רק עם מחצית מעובדי השלב השני בלבד. יתרה מזאת, הכתבות שניתנו לכל אחת מהקבוצות היו שונות לגמרי. כך מתאפשר לנו לבצע שני "קייפולי" אימון צולב (2-fold cross validation) מבלי לחשוש מזליגת מידע (data leak) בין נתוני האימון למבחן. על מנת להימנע משיפור שנובע מהיכרות מוקדמת של העובד עם המשימה, דאגנו כי אף לא אחד מהעובדים יטפל באותה משימה ספציפית יותר מפעם אחת. הערכת המודל התבססה על סמך דיוק חיזוי ביצועי הציוותים במדגם הקבוצה השנייה (במערכת ההמתנה, holdout).

2.4 חזרה על הניסוי

את הניסוי הרצנו בשני תסריטים שונים. בתסריט הראשון הרצנו את הניסוי ללא שימוש במידע לגבי ביצועי העובד ההיסטוריים. מצב זה מדמה פרויקט חדש שבו למעביד אין מידע על העובדים. בתסריט השני לקחנו בחשבון קיום של מידע על ביצועי העובד בעבר. בתסריט הזה, באופן ספציפי, הגדרנו את ציון הביצוע ההיסטורי של העובד כממוצע ציוני הביצועים (ערך S) של חמש המשימות הראשונות של כל

הניסוי התבצע באמצעות 1,000 עובדים מקוונים שניסו בפלטפורמת שוק העבודה המקוון Amazon Mechanical Turk. שאלנו כל אחד מהעובדים שבע שאלות על פרטיהם האישיים, על מנת להשתמש בהם כמשתנים מסבירים כחלק מווקטור המאפיינים של העובדים (X_1 או X_2) - גיל, רמת השכלה (Education level), תחום עיסוק (employment industry), תפקיד (Job function), מצב תעסוקתי (Employment status), רמת הכנסה (Household Income) ומדינה. שאלות אלו נבחרו מכיוון שאלה הנתונים שבהם הפלטפורמה מאפשרת למעסיקים לסנן בחירת עובדים. מלבד שאלת הגיל, כל השאלות היו שאלות בחירה על בסיס הנתונים המתאפשרים לבחור בפלטפורמה. העובדים שנבחרו לניסוי היו רק כאלה שמעל 95% מהעבודות הקודמות שלהם אושרו על ידי מעסיקים קודמים (HIT approval rate for all requesters' HITs), מודד מקובל לבחירת עובדים מהימנים).¹

כתבות החדשות הפיננסיות בניסוי התבססו על מערך הנתונים של (Geva and Zehavi, 2014) שכרו את הנתונים מאתרי חדשות פיננסיים מובילים. בחרנו באופן אקראי 160 כתבות, מתוך הכתבות בעלות כמות מילים הנמצא ב-60 אחוז האמצעי (מעל אחוזון 20 ומתחת לאחוזון 80) מתוך מערך הנתונים שהחוקרים אספו, ובממוצע כלל כ-2,000 מילים על מנת להימנע מהטיות הנובעות מדעות קדומות אפשריות וידע קודם פוטנציאלי של העובדים ביחס לחברות המסקרות, ביצענו תהליך אנונימיזציה ידני של כל שמות החברות, מוצריהם ועובדיהם, כפי שהופיעו בטקסט המקור. על מנת להפוך את המשימה למורכבת יותר, הסרנו את הכותרות ואת תקציר הכתבה, אם היו.

בנוסף על הכתבות, מערך הנתונים שהשתמשנו בו כולל ערך "הנכונות להשקעה" האמיתי (Ground truth). ערך זה נקבע על ידי שילוב חוות דעת של שני מומחים בעלי תואר MBA עם רקע במימון על סולם שבין +5 ל- (5-). אל ערך "הנכונות להשקעה" האמיתי השווינו את ערך "נכונות להשקעה" כפי שנקבע על ידי ציוות זוג עובדים מסוים ומדד הצלחת הציוות ("ציון הצוות", S) נקבע כערך המוחלט של

1 כל אחד מ-500 עובדי שלב א' הכין 20 סיכומים קצרים תמורת שכר של 3.6 דולר. כל אחד מ-500 עובדי שלב ב' קבע 20 ציוני סנטימנט תמורת שכר של 1.5 דולר.

Table 2: Summary of the Models' Performance using Mean Absolut Error (MAE). (Best Results Are Marked in Bold)

	Model Name	Both Workers	Data from Step A worker	Data from Step B worker	Naïve Model
Without History	Linear Regression	1.243	1.228	1.240	1.223
	Random Forest Regressor	1.221	1.219	1.221	
	SVM	1.222	1.221	1.226	
With History	Linear Regression	1.151	1.251	1.152	1.255
	Random Forest Regressor	1.129	1.239	1.130	
	SVM	1.170	1.245	1.157	

של אלגוריתמי החיזוי לעומת מצב שבו יש מידע רק לגבי אחד העובדים. את החיזוי בהינתן כל סט של מידע ביצענו פעמיים גם כן – פעם אחת כאשר סט המידע כולל מידע על הביצועים ההיסטוריים של כל עובד, ופעם שנייה כאשר הוא אינו כולל מידע זה. נוסף על כך, לצורך בקר, בנינו מודל בייסליין נאיבי שאינו מקבל וקטור מאפיינים עובדים כלל. מודל זה "חווה" את הצלחת הצוות בנתוני המבחן לפי ממוצע הצלחת כלל הצוותים בנתוני האימון בלבד.

מימשנו את מודלי הרגרסיה לחיזוי, פונקציית f , באמצעות שלושה מודלי רגרסיה לחיזוי ציון הצלחת הצוות. בחרנו במודלים פשוטים אך רובוסטיים – Linear Regression ו-SVM regressor. הסיבות לבחירה זו הן מחד להימנע מ-overfitting, ומאידך להימנע מהצורך של אופטימיזציה נרחבת של מודלים שעשויה "למסך" את האינפורמטיביות הבסיסית של הנתונים.

נזכיר כי מודל רגרסיה ליניארית הוא מודל המתאים למודל ליניארי עם מקדמים $w = (w_1, \dots, w_p)$ שתכליתו למזער את סכום השאריות (Residual) בין משתני המטרה שנצפו בנתוני האימון על מנת לייצר קירוב ליניארי לתצפיות בנתונים חדשים. Random Forest הוא מודל שמאמן מספר רב של עצי החלטה שרואים כל פעם מדגם Bootstrap שונה של נתוני האימון, ובוחן משתנים לכל פיצול בעץ מרשימה אקראית על ידי שכלול מודל העץ הפשוט (Ho, 1995). לעומת זאת, מודל SVM (Support Vector Machines) בוחר את גבול ההחלטה (Boundary decision) הרחב ביותר בין משתני המטרה (Drucker, 1997).

עובד. ציון זה נוסף לווקטור המאפיינים של האישיים של כל עובד (H_i) , וקטור זה ישמש בהמשך לצורך אימון מודלי החיזוי.

Table 1: Key notation

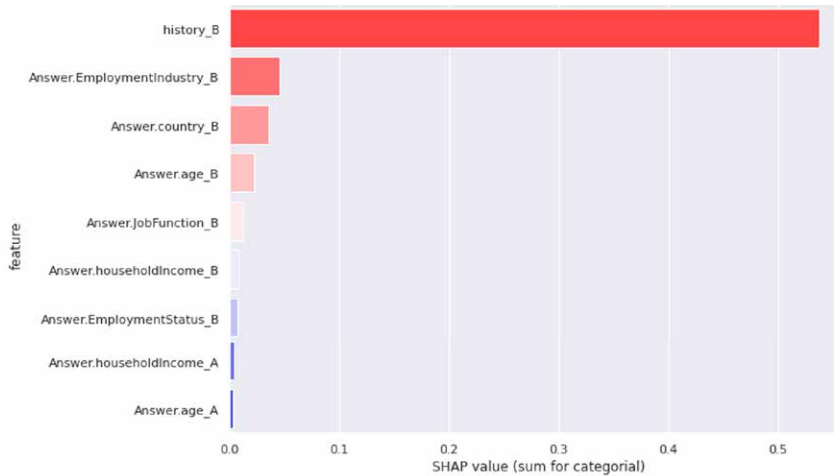
Notation	Description
$S_{i,j}$	The pair score of the worker i and j
T_i	Unique task i from type task T
$w_i^{(1)}$	Worker i 's for step A
x_I	Worker i 's features vector
H_i	The average history score of worker i
$f(w_i^{(1)}, w_j^{(2)})$	Model induced from worker i and j to predict S

3. ממצאים

סיכום התרחישים שנבדקו ומודלי החיזוי

כאמור, המחקר בחן את יכולת חיזוי ציון הצלחת הצוות, s , בהינתן סטים שונים של מידע. הבדיקה התבצעה מספר פעמים: פעם אחת כאשר סט המידע המשמש לחיזוי כולל וקטור מאפיינים שבו המאפיינים האישיים של שני העובדים, פעם שנייה כאשר וקטור המאפיינים כולל רק את מאפייני העובד הראשון, ופעם שלישית כאשר הוא כולל את וקטור המאפיינים של העובד השני בלבד (זה שמחליט בפועל על ציון הסנטימנט). כאמור, המטרה שלנו היא לבחון האם שילוב המידע משני העובדים תורם לתחזיות מדויקות יותר

איור 1: כיצד כל אחד מהמאפיינים השפיע על יכולת החיזוי של מודל המקבל נתונים של שני העובדים והביצועים ההיסטוריים שלהם



עם זאת, בהיעדר מידע היסטורי על העובדים (כתוצאה מ"התחלה קרה", Cold start), לא נתמכה השערותנו שניתן להשיג שיפור משמעותי למודל המלא ביחס למודל המסתמך על וקטור המאפיינים של אחד העובדים בלבד (ובחלק מהמקרים אף ממודל נאיבי הלוקח את ממוצע ציון הצוות מנתוני האימון).

על מנת לצמצם הטיות הנובעת מהתאמת יתר של מודלי החיזוי לנתונים הספציפיים שהוצגו לו ולמשימה הנבחנת, לא שונו פרמטרי המודלים מברירת המחדל, כפי שמופיע בספריות הקוד הרלוונטיות (Scikit-learn).

בחרנו למדוד את הצלחת מודלי הרגרסיה באמצעות ממוצע הטעות האבסולוטית (Mean Absolute Error), שהוא מדד מקובל בתחום לבעיות רגרסיה.

3.2 תובנות נוספות: ניתוח המאפיינים המסבירים ביותר

השתמשנו במודל SHAP (Lundberg et al., 2020) כדי לבחון מהם המאפיינים התורמים ביותר ליכולת חיזוי המודל. SHAP בודק את ממוצע התרומה השולית של כל אחד מהמשתנים לצמצום השגיאה של המודל באופן לוקלי (כיצד כל אחד מהמאפיינים עזרו לחזות – או לחלופין להסביר את הטעות מציון האמת – של דגימה אחת) וגלובלי (אילו משתנים עוזרים למודל לצמצם את השגיאה ופונקציית העלות, Loss function). המודל הוא אחד המודלים הפופולריים לפירוש תוצאות חיזוי (Interpretable AI).

המאפיינים התורמים ביותר ליכולת החיזוי לפי שיטה זו למודל הטוב ביותר (Random Forest לשני עובדים), כמתואר באיור 1, הם מאפייני עובדי שלב ב'. בפרט, המאפיין החוזה ביותר למודל המקבל נתונים היסטוריים הוא הביצועים ההיסטוריים של עובדי שלב ב'.

3.1 תוצאות הניסוי

בשורה התחתונה ניתן לומר כי ניצול מידע על שני עובדים משפר במעט את דיוק התחזיות ביחס למידע על עובד אחד, ככל שיש מידע מועיל בנתונים (היסטוריה של ביצועי עבר). באופן ספציפי, כפי שטבלה 2 מראה, בהינתן מידע היסטורי על ביצועי שני העובדים במשימות קודמות ואלגוריתם חיזוי מתאים (Random Forest), מודל שלמד ממאפייני שני העובדים שיפר את הדיוק ביותר מאחוז באופן מובהק ($p\text{-value} < 0.001$) לעומת המודל הטוב ביותר שלמד מאחד העובדים בלבד, או ממודל חיזוי נאיבי. המשמעות היא כי במצב שבו קיים מידע היסטורי על עובדים, נתמכה השערותנו ששימוש במידע על שני העובדים ישפר תחזיות ביחס לשימוש במידע של עובד אחד בלבד.

4. סיכום ומסקנות

העבודות המתבצעות במסגרת פלטפורמות גיוס העובדים המקוונים, לייצר שיתופי פעולה אפקטיביים, ולהציע צוותי עבודות למעסיקים פוטנציאליים בלי שהם יצטרכו לחפש עובדים. יתרה מכך, ככל שצוות יעבוד ביחד יותר זמן, כך ישתפרו ביצועיהם, והערך העסקי שהם נותנים למעסיק ולפלטפורמה יעלה. יצוין כי פלטפורמות גיוס העובדים באן-ליין שייכות לענקיות הטכנולוגיה, ולכן להם ככל הנראה מספיק מידע היסטורי על ביצועי העובדים על מנת להתגבר על בעיית "ההתחלה הקרה" הנובעת מהיעדר הידע.

במצבים שבהם אין נתונים היסטוריים על כל עובד, הסיבות היעדר הצלחתנו לשפר את החיזוי במקרים שיש בהם נתונים על שני העובדים, ביחס למקרים שבהם יש נתונים על עובד אחד בלבד, יכולות להיות מגוונות. ייתכן כי המשימה שבחרנו הייתה ספציפית מדי ונתנה משקל רב יותר להצלחת העובד השני. תיתכן גם קורלציה גבוהה בין הסנטימנט של העובד הראשון לכתבה (כפי שבא לידי ביטוי באופן כתיבתו את הסיכום) לסנטימנט שניתן על ידי העובד השני. עוד ייתכן כי היעדר שייכות שתי הקבוצות לאותה אוכלוסייה השפיעה על ההצלחה. בנוסף, בחירת אלגוריתמי החיזוי יכולה להשפיע על התוצאה, וייתכן שבחירת אלגוריתם אחר, בעל יכולות חזקות, היה משפיע על התוצאה.

על מנת לבחון הנחות אלו, להעלות את התוקף המחקרי ולשפר את ביצועי המודל, ניתן לבצע מחקרים נוספים. למשל, ניתן לבנות ניסוי נוסף הכולל משימה בעלת משקל שווה, פחות או יותר, לשני העובדים, לבחון אלגוריתמים נוספים של חיזוי, לבנות מודלים שונים לתתי-אוכלוסיות או לחלופין להרחבתן לצד שיפור החיזוי באמצעות התאמת המודלים לבעיה הספציפית.

ראשית, נציין כי לא הצלחנו להראות תמיכה גורפת בהשערותנו – מודל לחיזוי הצלחת ציוותי עובדים (בפלטפורמות שוקי העבודה המקוונים) הלומד ממאפייני שני עובדים אינו תמיד טוב יותר ממודל בסיסי הלומד ממאפייני אחד העובדים. עם זאת, תרומתה של עבודה זו עדיין כפולה – ניסחנו בעיית חיזוי הצלחת ציוותי עובדים כבעיה עסקית מוכוונת נתונים (Business data science problem). כמו כן, הצלחנו להראות שבמצב שקיימת היסטוריה קצרה (על ביצועי כל עובד בנפרד), חיזוי המבוסס על מאפייני שני עובדים ועל אלגוריתם חיזוי חזק מספיק (כמו Random Forest) ישיג ביצועים טובים יותר ביחס למודל הטוב ביותר המבוסס על מאפייני כל עובד בנפרד. הדבר מעיד ששילוב המידע של שני העובדים יכול לתרום ליכולת החיזוי של ביצועי צוותי עובדים.

השיפור שהושג במקרה זה אומנם היה קטן, אך ניתן למנות עבודות צוות טוריות ששיפור מה הוא בעל משמעות גדולה, דוגמת תיוג סרטונים לטובת מודלי למידת מכונה לרכב אוטונומי (כאשר העובד בשלב הראשון מסמן את האובייקטים והעובד בשלב השני מתייג אותם כפי שהתבצע ב-Camvid על ידי Gabriel, et al. (2008) ודומיו), הצבעה (שלב א') וזיהוי (שלב ב') של גידולים ומחלות לפי צילומי רנטגן (דוגמת CheXNet של Pranav, et al (2017)), שימושים ביטחוניים במודלי למידת מכונה (דוגמת שלב א' – זיהוי אדם ושלב ב' – זיהוי האם מסיג גבול) או אבטחת רשת (זיהו אנומליה בתעבורת רשת והחרגתה), וכתובת דוחות פיננסיים לחברות ציבוריות כפי ש-Fidler (2015) הציע.

גם להגדרת הבעיה יש השלכות עסקיות בעלות משמעות. בהינתן הגדרת הבעיה – יתכן שבעתיד עבודות נוספות יוכלו להציג מודלים משופרים שבאמצעותם ניתן לייצל את

tomergev@tauex.tau.ac.il

ד"ר תומר גבע

הגדרה פורמלית של בעיית המחקר

כאשר טעות החיזוי, $error$, משקללת את הטעויות בחיזוי $S_{i,j}$ עבור כל ציוות עובדים $i, j \neq i$ ($w_i^{(1)}, w_j^{(2)}$) על סמך מדד טעות מקובל כגון MAE.

חשוב לציין כי בהגדרה הנ"ל של הבעיה אנו מסתמכים לצורך החיזוי על וקטור המאפיינים אישיים של כל עובד, $x_i^{(1)}$ או $x_j^{(2)}$, הכולל רק תכונות ומאפיינים אישיים גלויים כגון מדינה, השכלה וכדומה. הדבר מתאים למצב שבו לא העסקנו את העובד בעבר ואין לנו נתונים לגבי ביצועיו במשימות דומות בעבר (בעיית Cold start).

תסריט שני:

בתסריט השני אנו נבחן הגדרה אלטרנטיבית של הבעיה, הכוללת גם נתונים ביצועיים היסטוריים של העובדים H_i, H_j , כאשר העובדים הועסקו במספר משימות קטן בעבר אצל אותו המעסיק. לצורך המחקר אנו בודקים תסריט שבו כל עובד הועסק בחמש משימות בעבר אצל אותו מעסיק. את היסטוריית ביצועי העובדים H נגדיר:

$$w_i^{(1)} \text{ עבור עובד } H_i^{(1)} = \frac{1}{\text{num tasks for } w_i^{(1)}} \sum_{j=1}^n S_{ij}$$

$$w_j^{(2)} \text{ עבור עובד } H_j^{(2)} = \frac{1}{\text{num tasks for } w_j^{(2)}} \sum_{i=1}^n S_{ij}$$

אנו נרצה לבחון האם מודל חיזוי המתבסס על המאפיינים האישיים של שני העובדים ונתוני ההיסטוריה שלהם, חוזה טוב יותר ממודל המסתמך על המאפיינים והנתונים ההיסטוריים של עובד אחד בלבד. כלומר:

$$error(f(x_i^{(1)}, H_i^{(1)}, x_j^{(2)}, H_j^{(2)})) < \min[error(f(x_i^{(1)}, H_i^{(1)})), error(f(x_j^{(2)}, H_j^{(2)}))]$$

בעיית המחקר שלנו עוסקת במצב שבו יש שני עובדים הנדרשים לבצע משימה טורית מסוג T. לדוגמה, משימה שכוללת שלב ראשון של סיכום מידע ושלב שני של מתן ציון לסנטימנט של המידע, או למשל שלב ראשון של תרגום מידע ושלב שני של סיכום. הניסוי עוצב כך שלעובד המועמד לבצע את השלב הראשון עבור המשימה הטורית יש וקטור תכונות הכולל סל תכונות פומביות הזמינות לסינון על ידי המעסיק בפלטפורמת העובדים המקוונים (דוגמת Amazon MTurk או פרוליפיק), כגון גיל, מדינה, הכנסה, השכלה וכדומה, לצד ציון ביצועי עבר. הגדרנו שותף זיווג פוטנציאלי – עובד עם וקטור תכונות אישיות כמועמד לבצע את השלב השני במשימה. מטרתנו היא לחזות ציון ביצועיים הדדי המשקף את הביצועיים המשותפים הצפויים במשימה T, אם נשייך את העובדים זה לזה באופן זמני לצורך ביצוע המשימה. דוגמא לציין אותו אנו הוצים לחזות, במשימה זו שלבית הכוללת סיכום מידע וקביעת הסנטימנט שלו – היא מידת הטעות בקביעת הסנטימנט.

תסריט ראשון:

במחקר זה ברצוננו לבחון את יכולת החיזוי של חיזוי $S_{i,j}$. באופן ספציפי ברצוננו לבחון האם מודל חיזוי המתבסס על מידע הכולל את וקטור המאפיינים ($x_i^{(1)}, x_j^{(2)}$) של שני העובדים המבצעים משימה טורית, יהיה מדויק יותר מחיזוי המתבסס על וקטור מאפיינים של כל אחד משני העובדים הללו. כאשר אנו משתמשים במודל חיזוי f (דוגמת Random Forest או גרסיה לינארית) המקבל כקלט וקטור מאפייני העובדים ונתונים אחרים ומחזיר בפלט את תחזית ביצועי הצוות. השערתנו היא שטעות חיזוי ביצועי הצוות של מודל המתבסס על נתוני שני העובדים, $error(f(x_i^{(1)}, x_j^{(2)}))$, קטנה יותר ממודל המתבסס על נתוני אחד העובדים בלבד. כלומר:

$$error(f(x_i^{(1)}, x_j^{(2)})) < \min[error(f(x_i^{(1)})), error(f(x_j^{(2)}))]$$

- Arias, Michael, Jorge Munoz-Gama, and Marcos Sepúlveda. "A multi-criteria approach for team recommendation." *International Conference on Business Process Management*. Springer, Cham, 2016.
- Brocco, Michele, Claudius Hauptmann, and Evi Andergassen-Soelva (2011). "Recommender system augmentation of HR databases for team recommendation." *22nd International Workshop on Database and Expert Systems Applications*. IEEE.
- Campion, Michael A., Gina J. Medsker, and A. Catherine Higgs (1993). "Relations between work group characteristics and effectiveness: Implications for designing effective work groups." *Personnel psychology*, 46 (4), 823-847.
- Codagnone, Cristiano, Fabienne Abadie, Federico Biagi (2016). "The Future of Work in the 'Sharing Economy'. Market Efficiency and Equitable Opportunities or Unfair Precarisation?" *Institute for Prospective Technological Studies, JRC Science for Policy Report EU*.
- Cramton, C. D. (2001). The mutual knowledge problem and its consequences for dispersed collaboration. *Organizational Science*, 12, 346–371.
- Datta, Anwitaman, Jackson Tan Teck Yong, and Anthony Ventresque (2011). "T-RecS: team recommendation system through expertise and cohesiveness." *Proceedings of the 20th international conference companion on World wide web*. ACM.
- Datta, Anwitaman, Jackson Tan Teck Yong, and Stefano Braghin (2014). "The zen of multidisciplinary team recommendation." *Journal of the Association for Information Science and Technology* 65 (12), 2518-2533.
- Demartini, Gianluca (2013). "Pick-A-Crowd: Tell me what you like, and I'll tell you what to do." CIDR.
- Dorn, Christoph, and Schahram Dustdar (2010). "Composing near-optimal expert teams: a trade-off between skills and connectivity." In OTM Confederated International Conferences: On the Move to Meaningful Internet Systems. *Springer Berlin Heidelberg*, 472-489.
- Driskell, James E., Eduardo Salas, and Tripp Driskell (2018). "Foundations of teamwork and collaboration." *American Psychologist* 73 (4), 334.
- Driskell, James E., Paul H. Radtke, and Eduardo Salas (2003). "Virtual teams: Effects of technological mediation on team performance." *Group Dynamics: Theory, Research, and Practice* 7 (4), 297.
- Drucker, Harris; Burges, Christ. C.; Kaufman, Linda; Smola, Alexander J.; and Vapnik, Vladimir N. (1997); "Support Vector Regression Machines", in *Advances in Neural Information Processing Systems* 9, NIPS 1996, 155–161, MIT Press.
- Fidler, Devon (2015). "Here's how managers can be replaced by software." *Harvard Business Review* 21.
- Geva, Tomer, Maytal Saar-Tsechansky, and Harel Lustiger (2019). "More for less: adaptive labeling payments in online labor markets." *Data Mining and Knowledge Discovery*, 1-49.
- Geva, Tomer, Maytal Saar-Tsechansky (2016). "Who's A Good Decision Maker? Data-Driven Expert Worker Ranking under Unobservable Quality". *In Proceedings of the International Conference on Information Systems (ICIS)*.

- Golshan, Behzad, Theodoros Lappas, and Evimaria Terzi (2014). "Profit-maximizing cluster hires." In Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 1196-1205.
- Greg Little, Lydia B. Chilton, Robert C. Miller, Max Goldman. (2009). *TurKit: Tools for iteratives on Mechanical Turk*. In Proceedings of the ACM SIGKDD Workshop on Human Computation (HCOMP '09), 29-30.
- Ho, Tin Kam (1995). "Random Decision Forests". Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14-16 pp. 278-282.
- Hoegl, Martin, K. Praveen Parboteeah, and Hans Georg Gemuenden (2003). "When teamwork really matters: task innovativeness as a moderator of the teamwork-performance relationship in software development projects." *Journal of Engineering and Technology Management* 20 (4), 281-302.
- Ipeirotis, Panagiotis G (2010). "Analyzing the amazon mechanical turk marketplace." *XRDS: Crossroads, The ACM Magazine for Students* 17 (2), 16-21.
- Ipeirotis, Panos G., Foster Provost, Victor S. Sheng, and Jing Wang (2014). "Repeated labeling using multiple noisy labellers". *Data Mining and Knowledge Discovery* 28 (2), 402-441.
- Kargar, Mehdi, Morteza Zihayat, and Aijun An (2013). "Finding affordable and collaborative teams from a network of experts." In *Proceedings of the SIAM International Conference on Data Mining (SDM)*, 587-595.
- Kittur, Aniket, et al. (2011) "Crowdforge: Crowdsourcing complex work." Proceedings of the 24th annual ACM symposium on User interface software and technology.
- Kittur, Anniket, J. V. Nickerson, M. Bernstein, E. Gerber, A. Shaw, J. Zimmerman, M. Lease, and J. Horton (2013). "The future of crowd work". In Proceedings of the 2013 Conference on Computer Supported Cooperative Work. ACM, 1301-1318.
- Kokkodis, Marios, and Panagiotis G. Ipeirotis (2015). "Reputation transferability in online labor markets." *Management Science* 62 (6), 1687-1706.
- Kuek, Siou Chew, Cecilia Paradi-Guilford, Toks Fayomi, Saori Imaizumi, Panos Ipeirotis, Patricia Pina, and Manpreet Singh (2015). "The global opportunity in online outsourcing." World Bank Report.
- Li, Hongwei, Bo Zhao, and Ariel Fuxman (2014). "The wisdom of minority: Discovering and targeting the right group of workers for crowdsourcing." Proceedings of the 23rd international conference on World wide web.
- Liemhetcharat, Somchaya, and Manuela Veloso (2014). "Weighted synergy graphs for effective team formation with heterogeneous ad hoc agents." *Artificial Intelligence* 208, 41-65.
- Lundberg, S.M., Erion, G., Chen, H. et al. (2020) "From local explanations to global understanding with explainable AI for trees" *Nat Mach Intell* 2, 56-67.
- Rajpurkar, Pranav, et al. (2017) "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning".
- Raykar, Vikas C., Shipeng Yu, Linda H. Zhao, Gerardo H. Valadez, Charles Florin, Luca Bogoni, and Linda Moy (2010). "Learning from crowds." *The Journal of Machine Learning Research* 1 (11), 1297-1322.
- Wang, Jing., Ipeirotis, Panagiotis. G., and Provost, Foster (2017). "Cost-Effective Quality Assurance in Crowd Labeling" *Information Systems Research*.