הקאתון IML – יוני 2023 – Hotel California

רקע ותחילת עבודה

התחלנו **בביסוס הידע שלנו על הדומיין** על סמך מחקרים אינטרנטיים וניתוח הדאטא הקיים והפיצ׳רים שנדגמו – שמותיהם, טיפוסיהם, יחידות המידה וכן בחינת סטטיסטיקות של פיצ׳רים שונים מהדאטא **שהפרדנו לסט האימון** (עד כמה השפיעו, קורלציה בין פיצ׳רים, ובעיקר מול ה-label) והסקנו מסקנות על איך להיעזר בפיצ׳רים הקיימים.

הבחנו שהדאטא נדגם על בסיס נתונים מיוני עד ספטמבר 2018, כלומר – **המודל צפוי לעבוד טוב יותר לתחזיות על עונת הקיץ וכן לא יצפה שינויים בלתי צפויים כמו למשל התפרצות הקורונה**.

השקענו מחשבה רבה בבחינת הפי׳צרים בביקורתיות וערכנו דגימות לא הגיוניות בסט האימון: בדקנו ערכים לא תקינים בדאטא של ה-train (הן בתיוג והן בפיצ׳רים)[1], הוספנו פיצ׳רים חדשים[2], יצרנו פיצ׳רים קטגוריאלים[3], הסרנו עמודות לא רלוונטיות ובחנו מה הפורמט של ערכים של כל פיצ׳ר - המרנו ערכים בוליאנים לבינארים על מנת לקבל ערך מספרי וכן החלפנו ערכים חסרים בערכים מתאימים[4].

שאלה 1 – Cancellation Prediction

באופן כללי – בחנו מודלים שונים על סט האימון, והרצנו אותם על סט הואלידציה על מנת לבחון את הביצועים שלהם (תוך וידוא שאנחנו לא מבצעים overfitting על סט האימון, ובחינת hyper-parameters שונים), במקביל בחנו איך הורדה והוספה של פיצ׳רים שונים משפיעה על שגיאת ההכללה.

בתחילה בחרנו baseline של מודל פשוט, ובחנו מודלים שונים שנלמדנו בהרצאות וכן מודלים שהכרנו לראשונה מחיפושים באינטרנט המשמשים בניתוחי דאטא דומים. לגבי כל מודל, שינינו את ההייפר-פרמטרים שלו בצורה הדרגתית על מנת לשחק עם המורכבות שלו ולנסות להגיע לנקודה די טובה מבחינת bias-variance tradeoff. כך, מצאנו למשל כי אדאבוסט כמטא-לרנר עם decision stumps בעומק 1 משיג תוצאה יחסית טובה עם כמאה לומדים. תוצאה טובה אף יותר התקבלה באמצעות שימוש במודל של Random forest המשתמש בכמאה לומדים. בסופו של דבר, בחרנו במודל מסוג xgBoost – זהו מודל המבוסס Gradient Boosting וגם הוא משתמש בועידה של לומדים פשוטים (עצי החלטה), ולהבנתנו אלו כלים איכותיים במיוחד עבור בעיות רגרסיה וקלסיפיקציה. בחרנו בכלי זה מכיוון שהוא הביא את התוצאה הגבוהה ביותר מבין שאר המודלים, ואת המודל הספציפי בדקנו באמצעות שינוי ערכים של הפרמטר learning_rate שנע בין 0 ל-1– ככל שהוא קרוב ל-0, נקבל מודל קשיח יותר, כלומר bias גבוה. מנגד, ככל שהוא קרוב ל-1 נקבל variance גבוה ונסתכן באוברפיטינג.

---

[1] למשל, וידאנו שכמויות ומחירים אינם שליליים, שהזמנים המתועדים הגיוניים (לדוג׳ שהזמנות התבצעו לפני זמן הצ׳קאין) וכן שהתיוגים הגיוניים (למשל – לא ייתכן ביטול של הזמנה שכבר התרחשה).

[2] למשל, הסרת כפילויות, זמן מאז שהאתר עלה לאוויר, האם ההזמנה בוצעה בשעות הלילה, זמן השהייה שהוזמן למקום, כמה ימים מראש ההזמנה התבצעה, האם ההזמנה תוכננה לסופ״ש, כמה בקשות מיוחדות היו בכל הזמנה, החלפנו את שמות המדינות שמהן בוצעו ההזמנות לערך ה-GDP המתאים לכל מדינה (או הממוצע אם לא הופיעו).

[3] סוג התשלום, אפשרויות החיוב

[4] למשל, החלפה בסט האימון של בקשות מיוחדות שלא מולאו באפסים, והוספת הערכים הממוצעים של הפיצ׳רים מסט האימון בתאים חסרים בסט הולידציה.

שאלה 2 – Cost of Cancellation

בשאלה זו ביצענו תחילה תהליך של preprocess הדומה לתהליך שביצענו בשאלה 1. אולם **ביצענו מספר התאמות ושינויים שנדרשו לשלב ה-preprocess הקודם – העיקרית מביניהן היא נירמול של הדאטא**, שכן על אף שב-preprocess הסרנו ערכים חריגים לכל פיצ׳ר, סדרי הגודל של הפיצ׳רים שונים ויפגעו בתהליך הרגולריזציה. דוגמה למשתנה שנרמלנו הוא ה-GDP של כל מדינה, שכן מדובר בערכים שעלולים להגיע לגדלים של מעל ל-20,000 ביחידות המידה שנמדדו (בחנו זאת הן על ידי ניתוח חוזר של סטטיסטיקות הפיצ׳רים והן בחיפוש אחר המונח באינטרנט). בנוסף, נרמלנו פיצ׳ר שמהווה את הזמן שעבר מאז עליית האתר לאינטרנט עד לזמן ה-checking, שכן גם הוא עלול להיות גבוה מאוד ולהחריג את גודל הנורמה הנמדדת בבעיות הרגרסיה כפי שלמדנו. בסטטיסטיקות שבחנו בשלב זה שמנו לב גם לערכים מה-preprocess הראשון שהקורלציה שלהם עלתה וכנראה תורמים יותר למודל, למשל – הפיצ׳ר שיצרנו שמעיד אם השהייה התרחשה בסופ״ש שהייתה ביחס ישר עם המחיר, וכן כמות ימי השהייה.

בשלב זה נדרשנו לתת הערכה למחיר שנפסיד בשל ביטולים ע״י תיוג של מחיר במידה שההזמנה בוטלה וערך של מינוס אחד אם ההזמנה לא בוטלה. ביצענו את ההערכה הזאת במספר שלבים:

1. גם בשלב זה **פיצלנו את הדאטא שקיבלנו לסט אימון ולסט ולידציה** (ומכל אחת מהקבוצות הפרדנו את העמודה של "original selling amount" כך שתשמש עבורנו בתור ה-y של כל אחת מהן).
2. **אימנו את סט האימון על מודלים שונים ובדקנו את טיב הפרדיקציה של כל אחד מהם באמצעות קבוצות הואלידציה** (כפי שעשינו בשאלה 1). בין המודלים שבחנו היו מודלים שנלמדו בכיתה כמו Polynomial Regression, Ridge, Linear Regression... בנוסף, בחנו גם מודלים חדשים שלמדנו עליהם מהאינטרנט. **לבסוף בחרנו ב-Lasso שהניב את השגיאה הנמוכה ביותר.** למודל זה אכן יש יתרונות במקרה זה משום שהוא מדייק את סינון הפיצ׳רים שלנו, שכן הוא נותן משקל שמאפס פיצ׳רים מסוימים, וכן הוא רגיש פחות ל-outliers.
3. **כמובן שתוך כדי עבודה למדנו אילו פיצ׳רים תורמים לפרדיקציה ואילו פחות רלוונטיים.** למשל: בסעיף הקודם התייחסנו למדד ה-GDP של המדינה שממנה מגיע הלקוח, אולם הבחנו שגודל זה זניח ביחס למדד ה-GDP של המדינה שבה יעד ההזמנה. על כן הסרנו את הפיצ׳ר הראשון והוספנו את הפיצ׳ר השני.
4. לאחר מכן איחדנו את עמודת המחירים שחזינו יחד עם מטריצת הדגימות ושלחנו אותה לקבלת פרדיקציה ע״י המודל שאימנו בשאלה 1 (על כל דאטא האימון שקיבלנו), שממנה **קיבלנו חיזוי לאילו הזמנות יבוטלו**.
5. על פי האינדקסים שהיו אפסים בחיזוי ביטול ההזמנות, שינינו את ערכי חיזוי הכספים שיצרנו כך שהפכנו את התאים שהתאימו לאינדקסים אלו לערך מינוס 1. **כמובן שבדקנו גם את תוצאות אלו עם ערכי הואלידציה שחילקנו ועם המודלים השונים שבחנו בתחילה.**