אמיר וינפלד | איתי ברכה
אביתר שאול | אביהו אלמוג

Hackathon Challenge 1: Hotel Cancellation

תיאור Dataset והאתגרים איתו:

- קיבלנו 58,659 רשומות עם 38 פיצ'רים עם משתנה שעניין אותנו בעיקר
- אחד האתגרים היה התמודדות עם מידע חסר. דרך ההתמודדות שלנו הייתה שונה בהתאם לאופי הפי'צר: את מרבית המשתנים הרציפים השלמנו באמצעות ממוצע של הערכים המלאים. עבור משתנים קטגוריים ביצענו השלמה באמצעות בחירת הערך השכיח ביותר (היסטוגרמה).
- עבודה עם משתנים קטגוריים: יש משתנים שעבורם ביצענו one hot encoding. אבל בשביל לחסוך גידול המרחבים, ניסינו לצמצם פיצ'רים (למשל במקום 7 משתני dummies על "יום בשבוע" בחרנו לעבוד עם ביטוי בוליאני של "האם סוף שבוע"
- נציין כי אחד הקשיים הגדול ביותר שלנו היה דף ההוראות, עם סתירות ביחס לתוכן המידע (ועם תוכן ישן מהאקאתונים של שנים שעברו). הדבר גרם לנו להמון זמן שבוזבז.

תהליך Preprocessing:

- כפי שהומלץ בהרצאות בשביל להימנע מזיהום כל ה Dataset. הדבר הראשון שעשינו היה לחלק את המידע שקיבלנו ל 4 קבוצות בגודל שווה (כל אחת 25%). נציין שלמעט רשומות כפולות, **לא** מחקנו דגימות בכלל.
- שמרנו מחוץ לסט האימון את ה id, שכן צריך אותו לפלט, אבל הוא לא מידע רלוונטי לתהליך הלמידה.
- עבור עמודות של תאריכים: הוספנו פיצ'ר "זמן (בשעות) מההזמנה עד Check-In", הוספנו עמודה של "משך השהות (בשעות)". את העמודה של "יום בשבוע" החלפנו ב "האם החופשה מתחילה בסוף שבוע" (כלומר במקום לעשות dummies על 7 אפשרויות החלפנו את זה לפיצ'ר אחד בינארי). בשביל להתמודד עם הציקליות של השנה, העברנו את תאריך הCheck-In לייצוג פולארי (באמצעות פונקציות טריגונומטריות). לסיום, יצרנו פיצ'ר של גיל המלון (לפי האתר).
- הורדנו שהופיעו אחוז נמוך מהמקרים בשביל שנוכל להסיק משהו על כלל המידע (למשל hotel id)
- בשביל להתמודד בהמשך עם NaN, שמרנו את הממוצע של הערכים (לקובץ שייטען במהלך ה test) בשביל שכאשר נבצע preprocess על ה test set נוכל להשלים אותו ממידע שכבר למדנו.
- ביצענו בשעות הראשונות תהליך של המרת מטבע על מנת שיהיה בשער אחיד, אבל לאחר מספר שעות התפרסמה הודעה כי ניתן להניח שהמטבע אחיד, לכן התעלמנו מסוג המטבע

במהלך העבודה התאמצנו להגיע ראשית אל פתרון מהיר, משם להתקדם הלאה ולהרחיב אותו כל הזמן.
את העבודה עשינו עם GitHub, עבודה עם branches שונים וניהול גרסאות לנוחות הכתיבה.
השתדלנו להשתמש בכמה שיותר מודלי למידה שראינו בכיתה.

שיטות שניסינו ותוצאות שקיבלנו:

כאמור, על מנת להגיע לתוצאות בסיסיות ככל האפשר, כבר בשעות הראשונות ביצענו preprocessing על מספר קטן מאוד של פיצ'רים, תוך שאנו משמיטים שורות בעייתיות (שורות עם מידע חסר), זאת על מנת לבנות את ה control-flow של תהליך בניית המודל. כאשר חלק אחר מהקבוצה המשיך לבצע preprocessing ולהוסיף לנו עוד פי'צרים, כבר באמצע העבודה על המשימה הראשונה, סיימנו לעבד את מרבית הפיצ'רים.

- במשימה הראשונה פיצלנו את המידע ל 80% אימון ו 20% בדיקה. ניסינו מגוון שלמדנו בשליש הראשון של הסמסטר:,Logistic Regression, Linear Regression, K-Nearest-Neighbors K-NN, SVC,LDA, QDA. עבור הערכת הצלחת המודל השתמשנו במטריקה F1 macro. לאחר מכן, הוספנו עוד מידע, אשר התווסף שוב בתצורה של 80% לסט האימון, ו20% סט בדיקה אשר מתוכם לקחנו 10% עבור validation set. בשלב זה הרצנו Random Forest למציאת המספר העצים האידיאלי תוך תשומת לב ל Bias-Variance tradeoff.
- במשימה השנייה כבר היה את מרבית ה data לאחר עיבוד מקדים. בדקנו שוב Random Forest ואת הלומדים ה"בסיסיים". ובסוף מצאנו ששימוש ב AdaBoost מביא שגיאה יחסית נמוכה. ביצענו הרצה איטרטיבית של שההיפר פרמטר עבור n = 14 מביא שגיאה נמוכה. בתהליך האימון יצרנו בעצמו את y_true שכן היה ניתן לחשב אותו מתוך הנתונים, כאשר ביצענו פרסור למדיניות הביטולים של המלונות וחישבנו את סכום העסקה של המלון עם הלקוח.

- במשימה השלישית סיננו את הפיצ'רים שיש להם קורלציה גבוהה מאוד עם פיצ'רים אחרים (ולכן הסרה שלהם לא תשפיע). עם ובלי dummeis, באופן טבעי, מצאנו קשר חזק בין מספר מאפיינים, למשל קשר חיובי בין number of adult לבין מספר החדרים שהוזמנו. לאחר מכן בדקנו מתאם פירסון בין הפיצ'רים שנותרו לבין התחזית.
- במשימה הרביעית

<u>מודל הלמידה הסופי:</u>

בהתאם לשאלה, בחרנו את המודל הכי מתאים לעבודה עם סט המידע שלנו, בשאלה הראשונה בחרנו Random Forest, ובשאלה השנייה בחרנו לעבוד עם AdaBoost. ביצענו שמירה של האובייקטים לאחר תהליך הלמידה לפרוייקט על מנת שנוכל לטעון אותם מוכנים ולחסוך בזמן ריצה שוב.

<u>תצפיות, שגיאת הכללה ומה שציפנו לקבל:</u>

בכל העבודה שלנו עם המודלים עבדנו עם train, test וגם עם validation, ניתן לראות כאן את f1 score כפונקציה של מספר הלומדים עבור random forest