האקתון ימל בראבק!!!

בפרויקט זה חקרנו דאטא בנושא סרטן השד. למדנו המון על הנושא, בעזרת האינטרנט ואנשי תוכן, וניסינו להפיק את המירב מהדאטא הנתון.
האתגרים המרכזיים שהיו לנו בפרויקט הם:

1. למידה של מספר לייבלים - במשימה הראשונה, לכל דגימה הדאטה יכולים להיות כמה לייבלים. הדבר מאוד הקשה על הבנת הלייבלים, החלטת הלמידה מהם והעבודה עם מספר לומדים בו זמנית.
2. למידה של דאטא בנושא לא מוכר - המון המושגים הרפואיים לא הכרנו ונדרשנו ללמוד ולחקור כדי להבין את המידע טוב יותר
3. עבודה עם דאטה בשפה חופשית - חלק מהדאטא לא היה בקטגוריות ספציפיות, נאלצנו לזקק מן המידע את המירב בהתאם.

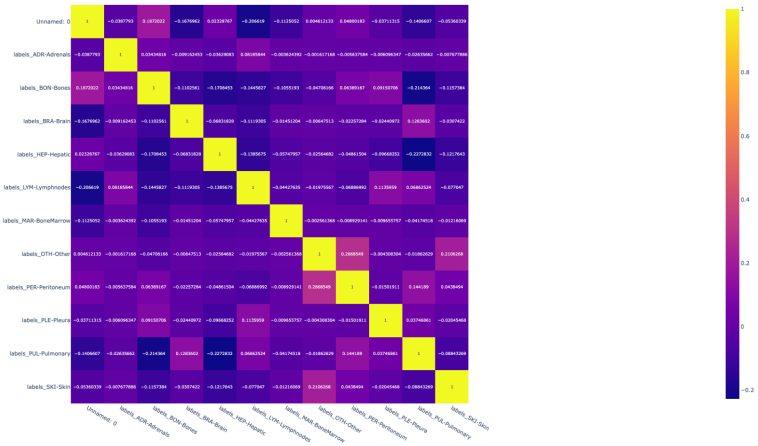
תהליך preprocess הוא החלק בו השקענו את רב זמננו. לכל פיצ'ר הבנו את הערכים האפשריים, למדנו מה ניתן להסיק על יחס ביניהם (יחס סדר או לא) ומאיזה מהדאטא לא כדאי ללמוד.

במהלך האקתון, חילקנו את הדאטה שיש ברשותנו לאימון ווולידציה. בהתחלה השתמשנו רק ב40% מהדאטא כאימון וב10% בולידציה. עם הזמן "פתחנו" לעצמנו עוד מידע לאימון עד שלבסוף הגענו ל70% אימון ו30% ולידציה. בכך חילקנו את הולידציה לכמה חלקים, כך יכולנו להסתכל באופן ביקורתי על הלמידה עד כה מבלי להסתכן ב Overfit.

ראינו שעיקר השיפור בתוצאה נבע מ preprocess איכותי יותר
בתחילת התהליך, הגענו לפרדיקציות מאוד טובות עבור סט הולידציה. התוצאות היו טובות מכדי להיות אמיתיות ועוררו חשד שאין הפרדה מוחלטת בין סט האימון לסט הולידציה. מהר מאוד הבנו שלכל פציינט יש כמה שורות בדאטה ואנו צריכים לוודא שכל השורות של אותו פציינט מופיעות באותו סט (אימון/ולידציה). כתוצאה מההפרדה החדשה קיבלנו תוצאות פחות טובות (משמעותית) וזה תאם את העובדה שעדיין לא עיבדנו טוב מספיק את הדאטה.

במשימה הראשונה ניסינו מספר לומדים שונים ולכולם תוצאות יחסית דומות. בחרנו בסופו של דבר להשתמש ב AdaBoost עם אלגוריתם בסיס - Decision Stump עם 60 איטרציות עבור כל אחד מהמקומות האפשריים לגרורה. אנו משערכים עבור כל דגימה האם יש גרורה בכל איזור בנפרד ומצמצמים את זה לבסוף לתשובה אחת מאוחדת. אחד האתגרים המשמעותיים היה להגדיר את הלייבלים בצורה משמעותית. האם מתייחסים לכל אזור בתור לייבל יחיד בלתי תלוי באחרים או שקבוצה של מספר לייבלים היא לייבל נפרד בפני עצמו.
חקירת פיצ'רים.

Labels correlation

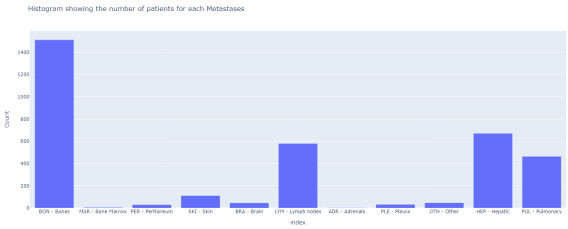


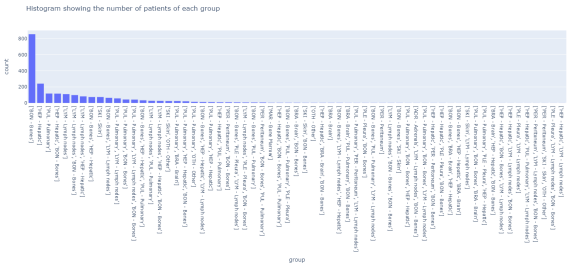
במשימה השנייה ניסנו לשערך את גודל הגידול בעזרת כל הנתונים. השתמשנו ברגרסיה לינארית עם אלגוריתם Ridge עבור רגולריזציה. מצאנו את פרמטר הרגולריזציה המיטבי בעזרת סריקה של ערכים והשווינו בין Ridge לLasso. ווידאנו שלא ביצענו Overfit על סט הולידציה בעזרת שינוי סט ולידציה לאחר קביעת הפרמטר.

כיוון חשיבה שניסינו לבדוק הוא האם קיים קשר בין אזורים שונים של גרורות. ההשערה שלנו הייתה למשל אם יש לנו גרורות באיזור מסויים בגוף נניח ברגל, אז אם יהיה עוד איזור נוסף שנפגע, אז שהוא יהיה קרוב אליו פיזית. לכן ניסינו לבדוק את הקורלציה בין הלייבלים שזה כיוון חשיבה אחר ממה שנהוג בדרך כלל. כפי שניתן לראות במפת החום המצורפת הקורלציה בין האזורים השונים חלשה (החזקה ביותר היא 0.3) ולכן זנחנו את כיוון החשיבה הזה.

Histogram

בדקנו את כמות הדגימות מכל קבוצת גרורות ובנוסף בדקנו את כמות החולים בכל גרורה.





אנו מצפים לתוצאה במשימה 1 במונחי f1 score:

Macro-0.1, Micro-0.3

ובמשימה 2:

MSE=2.5

התוצאות הצפויות נקבעו עפ"י טסטים שביצענו על מידע שלא נגענו בו עד לקראת הסוף שמדמה מידע חדש מהעולם האמיתי.

למדנו מידע רב בעזרת שיח עם איש מקצוע (רופא), לדוגמה תפקוד לקוי של בלוטת הלימפה הוא סימן מקדים חשוב להמצאות גידול סרטני והתפתחות גרורות. לכן ייצארנו עמודה נוספת הקשורה לקריטריון החמור ביותר של הבלוטה, כך שהמודל "נותן לה משקל כפול".