IML Hackathon

הדאטא והאתגרים בו:

הדאטא היה מאתגר ממספר בחינות, ראשית, איתרנו כפילויות רבות. עבור אותה חולה ואותו ביקור למעשה הופיעו מספר סמפלים שונים. כמו כן, היו חסרים נתונים רבים ואתגר משמעותי היה להחליט איך למלא את הפרטים החסרים.

ניקוי ופרה-פרוססינג:

בשלב הראשון, איחדנו את הכפילויות שהופיעו בדאטא, ומספר סמפלים שונים שהיו חלק מאותו ביקור אוחדו לאחד. לאחר מכן החלפנו משתנים קטגורים בone-hot והחלפנו תאים ריקים בערכים בעלי משמעות ואיחדנו מהויות דומות בפיצ'ארים מסויימים. בנוסף, התמודדנו עם עמודות שבהן היו ערכים שונים שהוקלדו ללא בקרה והפכנו אותן לבעלות סטנדרט אחיד. לאחר מכן יצרנו מספר פיצ'רים שונים בעלי משמעות כגון, יחס בין כמות קשרי לימפה חיוביים לבדיקה וקשרי לימפה שנבדקו, מספר ימים מהניתוח האחרון ופיצ'רים נוספים.

בניית המודל

ראשית, יצרנו מודל בייס-ליין על מנת שנוכל להשוות אליו מודלים מתקדמים יותר. עבור הלייבלים של מיקומי הגרורות, פיצלנו את העמודה לעמודה אחת עבור כל מיקום אפשרי (11 מיקומים אפשריים) של גרורה ואימנו מודל random-forest עם הפרמטרים הדיפולטים של sklearn עבור כל עמודה. הפרדיקציה הסופית היא איחוד של כל המיקומים שכל מודל חזה בנפרד.

בעבור המשימה השניה, חיזוי גודל גידול, השתמשנו כבייס-ליין ברגרסיה לינארית של sklearn עם פרמטרים דיפולטים.

לאחר מכן השתמשנו בבחירת פיצ'רים של sklearn וגילינו כי שימוש ב36 פיצ'רים הטובים ביותר נותן את התוצאות הטובות ביותר עבור רגרסיה.

לאחר מכן, הבנו כי השיטה הנאיבית של חיזוי כל לייבל בנפרד נותן תוצאות לא מספקות. מצאנו מודל הנקרא rakelD המסוגל להתמודדת עם בעיית מולטי-לייבל ומולטי-קלסיפיקציה. גילינו כי שימוש במודל זה משפר את התוצאות. בנוסף בחרנו, כי אם estimator אחד בלבד קובע כי יש גרורה אז בפרדיקציה הסופית נחזיר כי יש גרורה, זאת משום שעדיפה הטעות של חיזוי גרורה כאשר אין מאשר לחזות שאין גרורה כאשר יש.

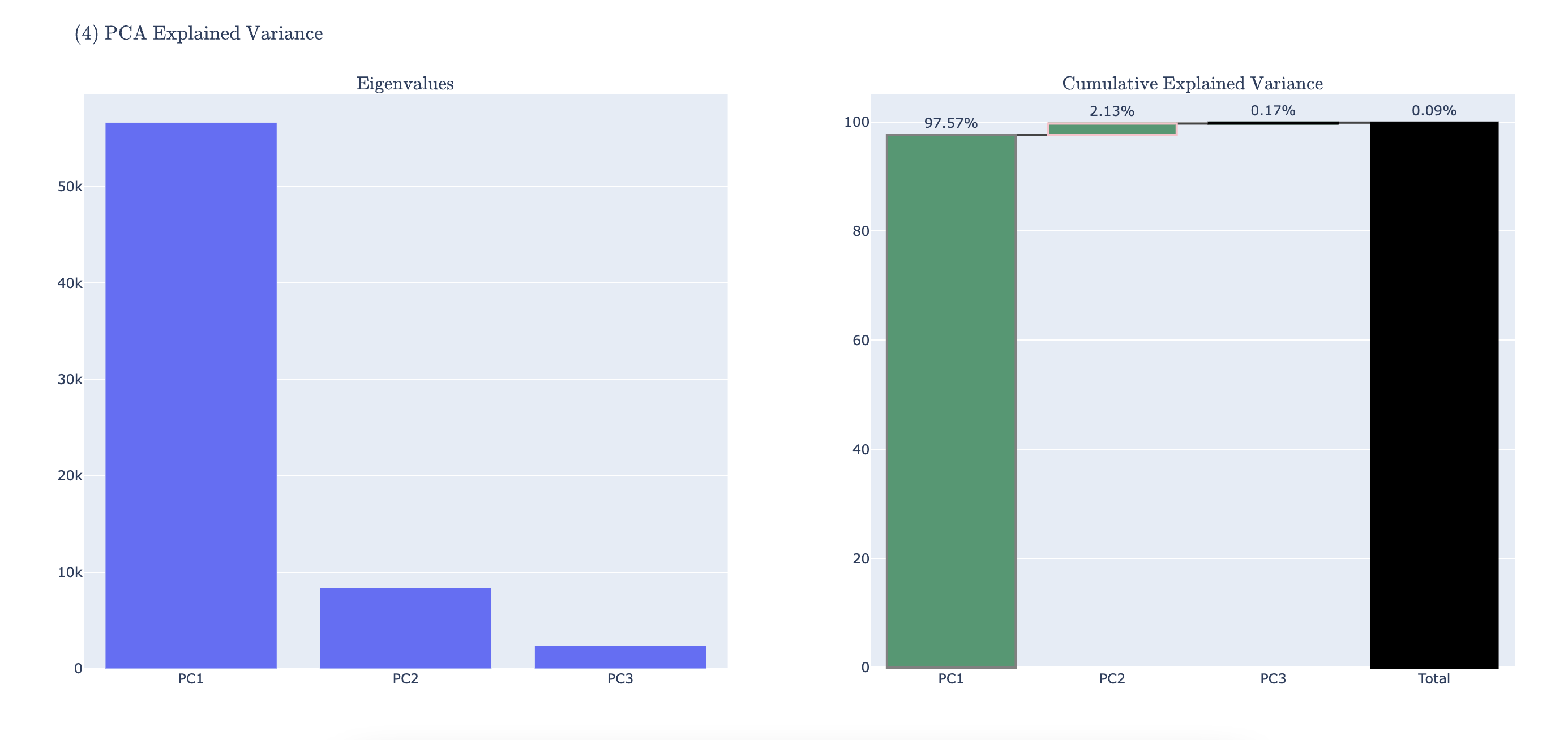
עבור כל אחד מהמודלים ניסינו להשתמש בשיטות נוספות לשיפור ביצועים, שימוש בהורדת מימד, קרנלים, אנסמבלים ועוד, ברוב המקרים גילינו כי שימוש בשיטות אלה לא שיפר את התוצאות ואף הוביל ל-over-fit.

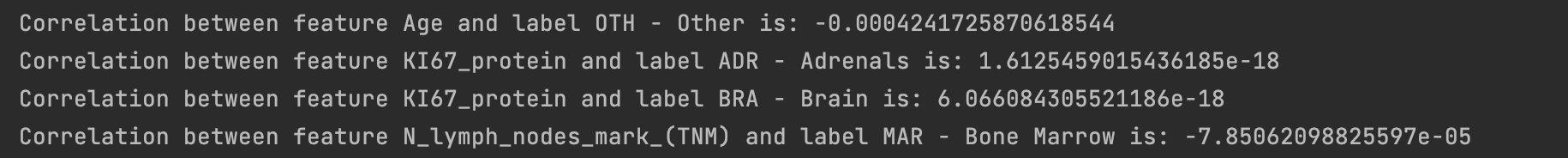
במשימה השנייה, של חיזוי גודל הגידול, מכיוון שמדובר ברגרסיה, קיבלנו גם גדלים שליליים. על כן ביצענו ReLU לערכים, כלומר לקחנו עבור כל ערך את המjtryום בין 0 לבין הערך.

לכל אחת מהמשימות, החלטנו לבצע comitee של מספר מודלים שונים, ולקחנו את הרוב/הממוצע (בהתאמה לסוג החיזוי הרצוי).

בעבור בעיית הרגרסיה, בסופו של דבר גילינו כי random-forest נותן את הביצועים הטובים ביותר ולכן בחרנו בו, בדקנו הייפר-פרמטרים שונים עד שמצאנו את הפרמטרים שעובדים בצורה המיטבית.

במהלך ניתוח הפיצ׳רים ניסינו להבין מי הם הפיצ׳רים הקורלטיביים ביותר, עשינו את הניתוח הזה בשתי דרכים

1) ויזואלית

2) בדיקת קורלציה באמצעות corr = cov / std\_feature / std\_y כפי שעשינו בתרגיל של ה- house price prediction