סעיף 1.2.3

בסעיף זה, נתבקשנו למצוא קבוצת פיצ'רים, להבנתנו קטנה מספיק שתהווה קבוצה איכותית דיה לאימון מודל שיחזה את ביטולי ההזמנות. ראשית הגדרנו כי קבוצה קטנה כלומר 5 פיצ'רים ואיכותית דיה כלומר היא משיגה תוצאות דומות או טובות מאלה שנתקבלו בסעיף 1.2.1.

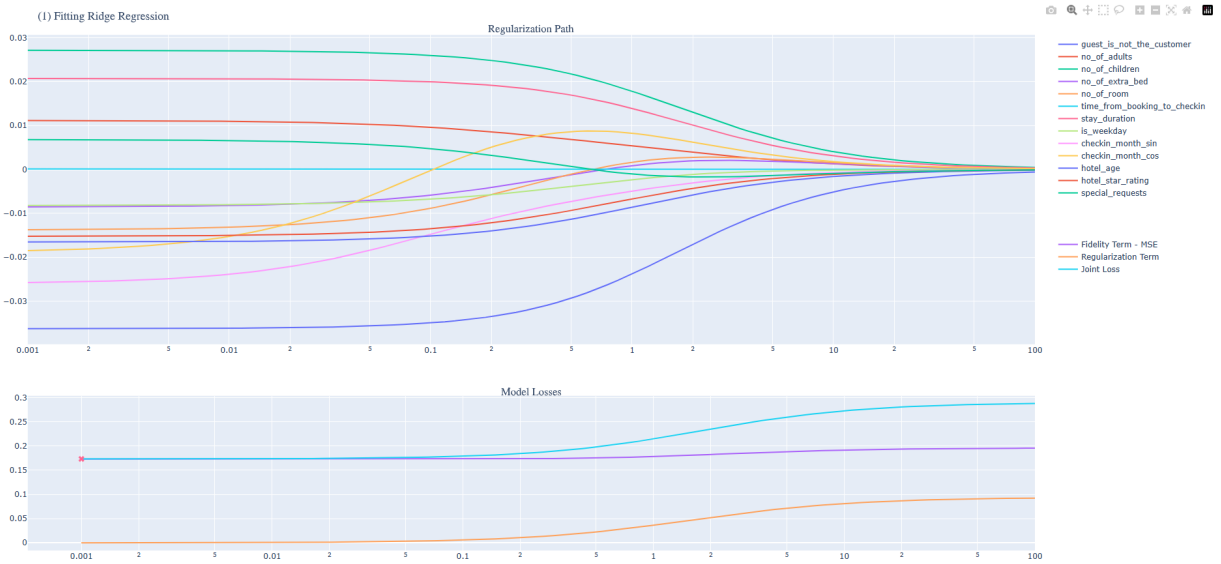
לשם כך, ההבנה הברורה שלנו הייתה, לשכוח מהפיצ'רים שיצרנו בדרך, כלומר כל ה dummies שנוספו במהלך הדרך שמרביתם בינאריים, על פי הגיון כללי וידע עולם, לא סביר שיכילו מספיק מידע כדי להתברג לרשימה היוקרתית. בנוסף, פסלנו את כל הפיצ'רים שעסקו ב- id (אלה של הלקוח ושל המלון ושל העיר והאזור) היות שמכל אחד מהם כשלעצמו כמעט ולא ניתן ללמוד דבר לדוגמא, קיימת כפילויות במידע שניתן לנו בקידוד הערים (ככל הנראה מתוך ההבנה שקידוד העיר רלוונטי רק בהתחשב בקידוד המדינה ואלה כבר 40% מהפיצ'רים - דבר שאיננו סביר לדעתנו).
בנוסף החלטנו לוותר על פיצ'רים שעסקו בשפות שכן, שפה איננה חד ערכית לאזור עולמי (אוסטרליה בריטניה וארה"ב שלושתן ביבשות שונות), שבהקשר זה (תחת ההנחה שהמילה 'ביטול' קיימת בכל שפה או ברובן המכריע) אין בה די ללמדנו כדי להכניס לרשימה.

כאשר נותרנו להכריע על בחירתם של 3 פיצ'רים מתוך 13 פיצ'רים (או 41 פיצ'רים עם ה-dummies), עברנו לסינון פיצ'רים באמצעות דרכים מתמטיות וכלים שנלמדו בקורס למשל, ביצענו בדיקות קורלציה בין כל הפיצ'רים שנותרו לעצמם (בזוגות) שהרי ברור שאם בכוונתנו 'לנצל' כמה שיותר מהמידע שניתן לנו, לא נרצה שתהא קורלציה גבוה בין הפיצ'רים שנכנסו לרשימה. בנוסף, בדקנו קורלציה בין פיצ'רים לביטול ההזמנה (הלייבל שלנו) על ידי שימוש *במתאם פירסון*, ומיון על פי הערכים המתקבלים ממנו. בין היתר גילינו כי יש קורלציה גבוהה יחסית חיובית עבור ה-dummie של "charge_option_Pay Later".

בנוסף, כהחלטה ערכית החלטנו שלא לייצר למען הסעיף פיצ'ר חדש מעבר למה שכבר נוצר במהלך התרגיל שהרי יש בכך טעם לפגם שכן, אמנם נעמוד במספר הפיצ'רים הרלוונטי, אך בעצם נוכל להשתמש מאחורה בהרבה יותר. ולתחושתנו לא הייתה זו הכוונה היחידה.
וכך בחרנו את הפיצ'רים הרלוונטיים, שהרי לדעתנו הם:
"charge_option_Pay Later", "time_from_booking_to_checkin", "stay_duration"

חדי העין יבחינו ש- charge_option_pay_later הינו אכן dummy (למרות שכמעט באופן ברור היה ניתן להתייחס אליו כמשתנה בינארי, משיקולים של אימון המודל הוא אכן לא הפך לכזה בסופו של דבר והסיבה שהוא נכנס לרשימה הייתה הקורלציה הגבוהה שנרשמה בינינו ללייבל, שעמדה על 0.374)

בנוסף, בחרנו להתבונן ברגוליזציה עבור עם Ridge Regression כדי לראות האם יש משהו מעניין שנוכל להסיק בעזרת האיור הבא לגבי הפיצרים.



לפי אופן ההבנה שלנו, הבחנו כי באיזור המשמעותי יחסית של ה-Model Loss, ניתן לראות כי רוב הערכים הולכים ומתכנסים לערך 0 באופן יחסית דומה זה לזה, ולכן בחרנו שלא להסיק מהאיור מסקנה קונקרטית לגבי אילו פיצ'רים משמעותיים יותר מאחרים.