

פרויקט סדרות עתיות וחיזוי

מזג אוויר דלהי 2013-2017

דאטה: <https://www.kaggle.com/datasets/sumanthvrao/daily-climate-time-series-data?resource=download>

תאריך הגשה: 06.04.2025

מגישות:

אליזה קבלקין - טכניון

חלא עבד-חלים - טכניון

חלק 1

הפרויקט שלנו מתמקד בניתוח נתוני מזג אוויר יומיים עבור דלהי, אחת הערים הצפופות והמשמעותיות בהודו [1], המאופיינת בשינויים אקלימיים קיצוניים לאורך השנה. מאגר הנתונים מכיל נתונים יומיים על תנאי מזג האוויר בעיר דלהי (הודו) ([Delhi data](#)), כולל המדדים הבאים:

- טמפרטורה ממוצעת (mean temp)
- לחות יחסית (humidity)
- מהירות רוח (wind speed)
- לחץ אוויר ממוצע (mean pressure)

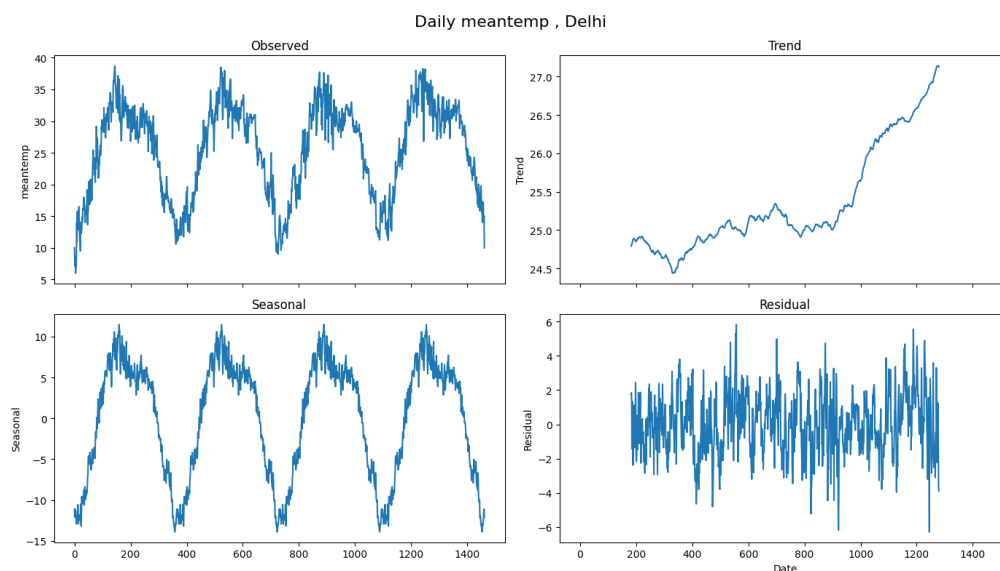
הנתונים נפרסים על פני תקופה של 4 שנים מ-1 בינואר 2013 ועד 31 בדצמבר 2016, ומכילים 1,462 רשומות.

מזג האוויר בדלהי משתנה באופן חזק לפי עונות השנה. [2] בקיצ, שמתחיל במרץ ונמשך עד יוני, חם מאוד עם טמפרטורות שמגיעות לעיתים קרובות מעל 40 מעלות צלזיוס, ולעיתים יש גם חום קיצוני עם טמפרטורות גבוהות במיוחד, תקופת המונסון (Wet season), בין יולי לספטמבר, מביאה עמה גשמים כבדים ולחות גבוהה, הסתיו (ספטמבר-נובמבר) והחורף (דצמבר-פברואר) הם הימים הנעימים ביותר, עם טמפרטורות נמוכות יותר, במיוחד בלילות החורף, שיכולות להגיע עד 5 מעלות צלזיוס.

פירוק המשתנים - בפירוק סדרה עתית, מטרתנו היא להבין את הגורמים השונים המשפיעים על הנתונים, כמו עונתיות, מגמות ושאריות (משפיעים אחרים).

הפירוק מאפשר לנו לבודד את השפעת העונתיות ולהעריך עד כמה היא מסבירה את השינויים בסדרה לאורך הזמן, כאשר אנו מבצעים את הפירוק, אנו מחפשים שהעונתיות תסביר באופן משמעותי את השינויים בעונות השונות, כך שהשאריות יהיו קטנות, אם השאריות נשארות גדולות לאחר הפירוק, זה מעיד על כך שהעונתיות אינה הגורם המרכזי בהסבר הנתונים, וייתכן שיש גורמים נוספים, כמו מגמות כלליות או רעש, שמסבירים את השינויים בצורה טובה יותר.

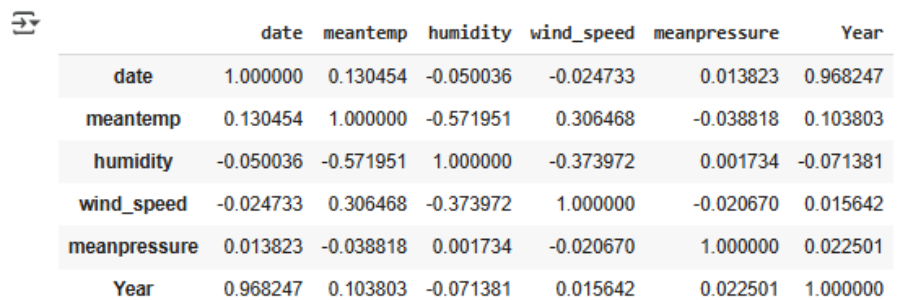
במקרה שלנו, ניתוח הפירוק מראה שהטמפרטורה הממוצעת הוא הפרמטר שמציג דפוסים עונתיים ברורים ומגמות משמעותיות ומסבירים היטב ביחס לשאריות לעומת הפרמטרים האחרים, ולכן הוא מהווה משתנה מרכזי שמוסבר על ידי עונתיות ומגמה, לכן נתייחס למשתנה של טמפרטורה ממוצעת כמשתנה מרכזי מהנתונים שיש לנו.



גרף העונתיות מחזק את ההבנה שמזג האוויר בדלהי נשלט על ידי מחזוריות עונתית ברורה, והוא יכול לשמש כבסיס לחיזוי טמפרטורות עתידיות לצרכים כמו תכנון תחבורה, צריכת אנרגיה וניהול משאבי מים. ניתן לראות זאת על ידי כך שטווח הערכים של העונתיות גדול מטווח הערכים של השאריות.

שלב אחרי זה הסתכלנו על טבלת קורלציה (Correlation Matrix) כדי לבחון את הקשר בין המשתנים השונים לבין הטמפרטורה הממוצעת ("meantemp").

מטרת הניתוח היא לזהות אילו משתנים – לחות יחסית (humidity), מהירות רוח (wind speed) ולחץ אוויר ממוצע (mean pressure) – קשורים באופן משמעותי לטמפרטורה.



	date	meantemp	humidity	wind_speed	meanpressure	Year
date	1.000000	0.130454	-0.050036	-0.024733	0.013823	0.968247
meantemp	0.130454	1.000000	-0.571951	0.306468	-0.038818	0.103803
humidity	-0.050036	-0.571951	1.000000	-0.373972	0.001734	-0.071381
wind_speed	-0.024733	0.306468	-0.373972	1.000000	-0.020670	0.015642
meanpressure	0.013823	-0.038818	0.001734	-0.020670	1.000000	0.022501
Year	0.968247	0.103803	-0.071381	0.015642	0.022501	1.000000

באמצעות הקורלציה, ניתן להבין האם יש קשר חזק, חלש או כלל לא קיים בין כל אחד מהמשתנים לטמפרטורה, מהי מגמת הקשר (חיובי או שלילי), מידע זה חיוני לחיזוי, שכן משתנים עם קורלציה גבוהה (בערך המוחלט) עם הטמפרטורה יכולים לשמש כ-גורמים חשובים במודל חיזוי. (העמודה / השורה של השנה לא רלוונטית, כי היא לא מייצגת פרמטר)

לאחר ניתוח טבלת הקורלציה, ניתן לראות כי הלחות היחסית (humidity) ומהירות הרוח (wind speed) הם המשתנים המשפיעים ביותר על הטמפרטורה הממוצעת. הקורלציה של הטמפרטורה והלחות הינה שלילית גבוהה יחסית (-0.57), כלומר ככל שהלחות עולה הטמפרטורה נוטה לרדת. לעומת זאת, עם מהירות הרוח הקורלציה חיובית בינונית (0.31), מה שמעיד על כך שכאשר הרוח מתחזקת, הטמפרטורה הממוצעת עשויה לעלות.

בהתבסס על התוצאות, נבחר להשתמש ב-humidity ו-wind speed כמשתנים מסבירים (exogenous parameters) במודל החיזוי שלנו.

חלק 2:

בחרנו להשתמש בשלושה מודלים שונים, SARIMA, PROPHET, HOLT-WINTERS על מנת לענות על השאלה של חיזוי המעלות בחצי השנה לאחר הדאטה. כדי להשוואות בין שלושתם ולבחור את המוצלח ביותר המתודולוגיה שבחרנו היא השוואה לפי הממוצע והשונות של הפרש החיזוי לתוצאות האמיתיות.

עבור SARIMA הבנו כי יש צורך בדיפרנציאציה היות והסדרה לא סטציונרית הסדרה המקורית תקרא סדרה א'.

בדקנו סטציונריות בדרכים הבאות:

- הראשונה היא בעזרת פירוק הסדרה לפי חלק 1 שבה רואים כי ישנה מגמה עולה עבור מדד הטמפרטורה.
- השנייה היא בעזרת מבחן ADF. אותו ביצענו על כל אחת מהסדרות כאשר רק עבור סדרה א' המבחן הראה כי אינה סטציונרית והשאר כן.

לגבי דיפרנציאציה החלטנו לבדוק הפרש בין יומיים עוקבים (סדרה ב'), הפרש בין יומיים במחזוריות של חודש (סדרה ג'), ובין ימים עם הפרש של שנה (סדרה ד'). הסיבה לבדיקה של יומיים בין חודשים מגיעה מכך כי מחקרים שונים מראים כי אנומליות ביחס לממוצע מתפלגות נורמלית [4].

לכל אחת מהסדרות הנ"ל גם הסתכלנו על ה-ACF וה-PACF כדי לאבחן אילו סדרי SARIMA מתאימים לנו.

בסדרה א' ראינו כי ה-ACF הוא מונוטוני יורד אך מאוד לאט וכי כ-60 הערכים הראשונים לפני הינם משמעותיים עבור יום ספציפי. זהו דבר שכן צפוי היות ומזג אוויר תלוי בפרמטרים שונים כמו מיקום כדור הארץ, השמש וכו' אשר משתנים באיטיות ומייצרים תלות בין המדידות. לפי ה-PACF אנחנו רואים קשר חזק בלאג הראשון, כמעט 1 ואחריו הלאגים לא שואפים לאפס אלא כל 10 לאגים ישנו לאג משמעותי מעט.

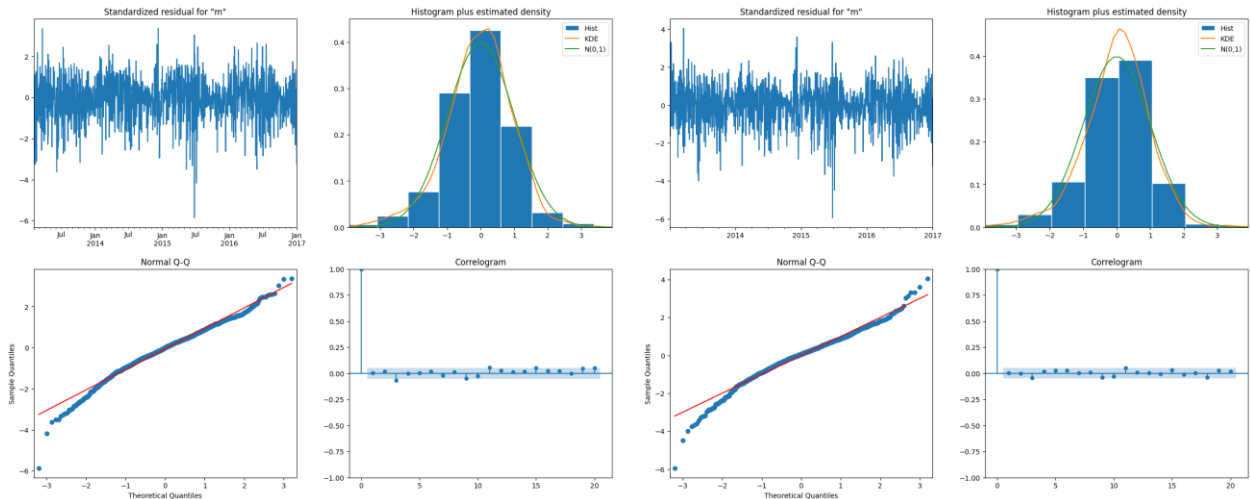
בסדרה ב' אנחנו רואים כי חלק משמעותי של הטרנד נעלם לפי ה-ACF וניתן לראות כי שני המדדים שואפים לאפס, דבר שמראה על סטציונריות ועל הפיכות. מהתבוננות בגרף ה-ACF ניתן לראות כי ה-3 לאגים הראשונים משמעותיים ולפי ה-PACF ה-4-5 לאגים משמעותיים מה שמכוון לבדוק סדרות כמו (ARIMA(4,1,3), MA(3), AR(4), ARIMA(4,1,3) מבין אלו. בסדרה ג', רואים כי ה-ACF יורד אך יש קפיצה ב-30 וה-PACF שואף לאפס גם כן עם קפיצות ב-30 וב-60. הדבר מצביע על כך כי ישנה MA עונתית AR עונתי אפשריים. לפי ה-PACF נראה כי ה-10 לאגים הראשונים משמעותיים גם. מכאן רצינו לבדוק את המודלים (SARIMA(10,1,1), (1,1,1,30) ו-(SARIMA(1,1,1,30).

בסדרה ד', רואים כי ה-ACF שואפים לאפס וכי ה-7 לאגים הראשונים הם משמעותיים. ה-PACF גם שואף לאפס וכי הלאג הראשון והרביעי משמעותיים. מסיבות של חישוב לא הרצנו את SARIMA עם $q=4$ מפאת זמן ריצה ויכולת חישוב של המחשב. נדגיש כאן כי היות וגודל הסדרה הוא 1462 לא ביצענו אגריגציה לשבועות או חודשים, כל אגריגציה כזו הייתה מורידה את גודל הסדרה מתחת לסף המינימלי הנדרש של 300 רשומות. דבר זה הוביל לכך שהמודלים המבוססים על סדרה ד' רצו זמן רב (לפחות שעה חלקם). בעקבות בעיית הזמן, דבר שהיו ידוע ומוכר בספרייה של פייתון [5] עבור עונתיות גדולה של 365, נותר על מודלים אלו בבדיקה.

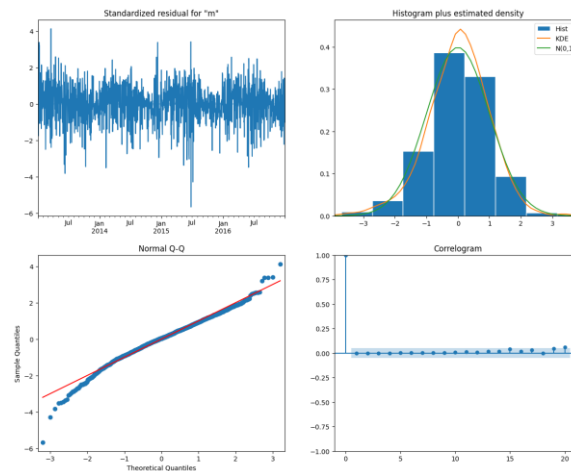
עבור כל מודל הדפסנו את התוצאות שלו וגם את הסיכום שמראה את השאריות. המודל המוצלח ביותר לפי ה-ACF היה מודל $(4,1,3)$ ARIMA. לפי בדיקת השאריות שלו ראינו כי המודלים של $(10,1,1)$ SARIMA ו- $(1,1,1,30)$ SARIMA היו משמעותיים יותר טובים בקרבתם להתפלגות נורמלית. ההבדל לפי שני המדדים השונים הוביל אותנו לבדוק את שלושתם המודלים:

SARIMA(1,1,1)(1,1,1,30)

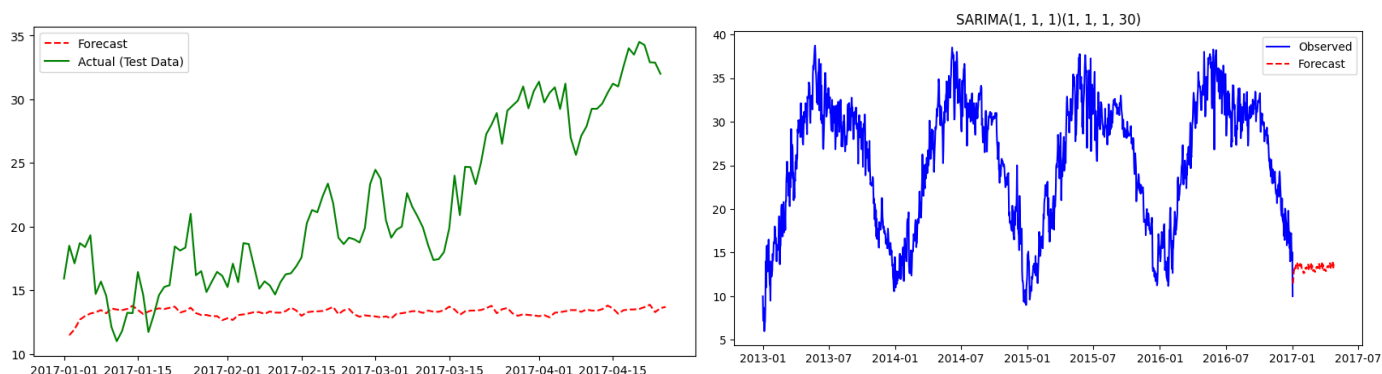
ARIMA(4,1,3)



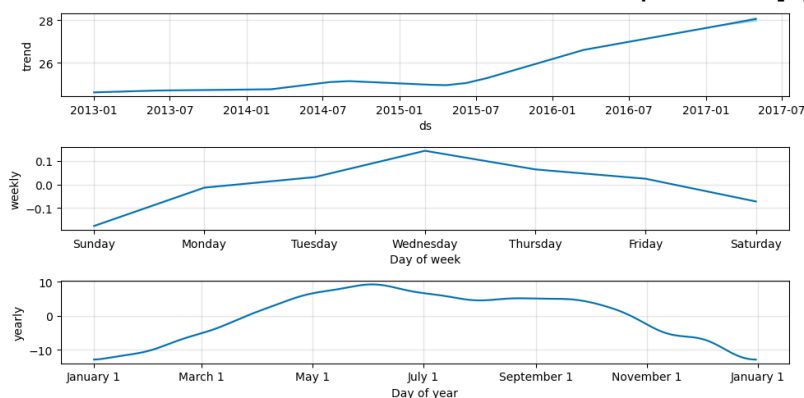
SARIMA(10,1,1)(1,1,1,30)



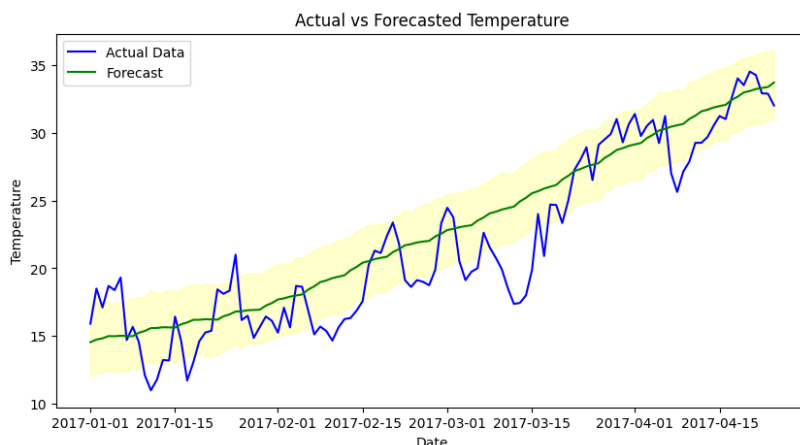
לאחר בחירת המודל ביצענו פרדיקציה ל-114 ימים הבאים והשוואנו לתוצאות סדרת ה-TEST שהייתה נתונה בקובץ נפרד. קיבלנו כי עבור $(4,1,3)$ ARIMA החיזוי היה בעייתי והראה מגמה יורדת חלקה כאשר התוצאות האמיתיות הראו עלייה. סיבה אפשרית לכך היא כי המודל הזה אינו ללא עונתיות והתחזית התבססה בעיקר על סוף שנת 2016 כאשר הטמפרטורות יורדות לקראת החורף. עבור $(10,1,1)$ SARIMA ו- $(1,1,1,30)$ SARIMA קיבלנו כי התחזית הייתה בקו מגמה ישר אך כן זיהה שינויים בין ימים ולא הראה מגמה חלקה. הדבר קורה כי ככל הנראה למרות ההתבוננות בהפרשים בין ימים המחזור אינו מספיק גדול, במאמר [4] החלוקה של הנתונים שלהם הייתה ל-3 חודשים (90 ימים ולא 30). מבין שלושת המודלים הללו המוצלח ביותר היה $(1,1,1,30)$ SARIMA לפי ממוצע השגיאה בין הערכים האמיתיים לחיזוי, MSE, והשונות שלה, 5.85, 72.9, 38.6 בהתאמה. היה דימיון גדול בין ערכיו לבין $(10,1,1)$ SARIMA ו- $(1,1,1,30)$ SARIMA לכן הבחירה היא גם על בסיס מספר הפרמטרים הנדרש למודל.



עבור מודל **PROPHET** לא עשינו דיפרנציאציות, ובחרנו להוסיף מחזוריות שנתית כפי שראינו בחלקים הקודמים אך גם שבועי כי ראינו כי בלעדינו נוצרת מגמה חזרה אשר מתקשה לזהות עליות וירידות לאורך העלייה הכללית. המודל עצמו זיהה מחזוריות לאורך השנה אשר דומה למה שראינו בפירוק והראה כי ישנה מגמת עלייה לאורך השנים, דבר שניתן להסביר על ידי ההתחממות הגלובלית והעלייה בטמפרטורה העולמית שם כן [3], כפי שניתן לראות בגרפים הנ"ל:



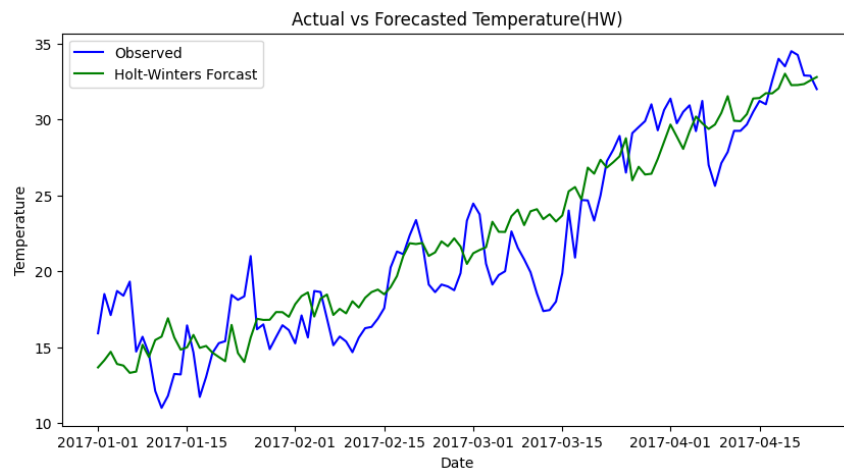
מבדיקת החיזוי ראינו כי המודל הצליח ללמוד כי ישנה מגמה עולה וניתן לומר כי גרף התחזית אשר יצר הוא חלק יותר בהשוואה לגרף התחזית של SARIMA, כלומר הוא לא מצליח לזהות פיקים בימים ספציפיים למרות תוספת הלמידה של המחזוריות השבועית.



בבדיקת ההפרש בין החיזוי לאמיתי ראינו כי המודל עבד בצורה טובה יותר מבחינת ממוצע והן מבחינת השונות היות והם היו קטנים יותר, ממוצע הפרש של -1.17, MSE של 7.47 ושונות של 6.1.

עבור מודל HOLT-WINTER לא היה צורך בשינויים של הסדרה וטיוב שלה. למודל הכנסנו התחשבות בטרנד ובעונתיות לפי הפירוק של חלק 1, והוספנו מחזור של 365 היות והדאטה שלנו הוא יומי לאורך כמה שנים.

מבחינת החיזוי הגרף נראה כי הצליח לזהות את המגמה וגם כן את הפיקים אך לא בעוצמה האמיתית שהייתה. הממוצע שלו עבור הפרש השגיאה היה -0.48, MSE של 7.07 ושונות של 6.83.



מבין שלושת המודלים, הממוצע הקטן ביותר (בערך מוחלט) והשונות הקטנה ביותר התקבלו במודל Holt Winters.

Model	Mean	Var	MSE
Sarima	5.85	72.9	38.6
Prophet	-1.17	7.47	6.1
Holt Winters	-0.48	7.07	6.83

חלק 3 :

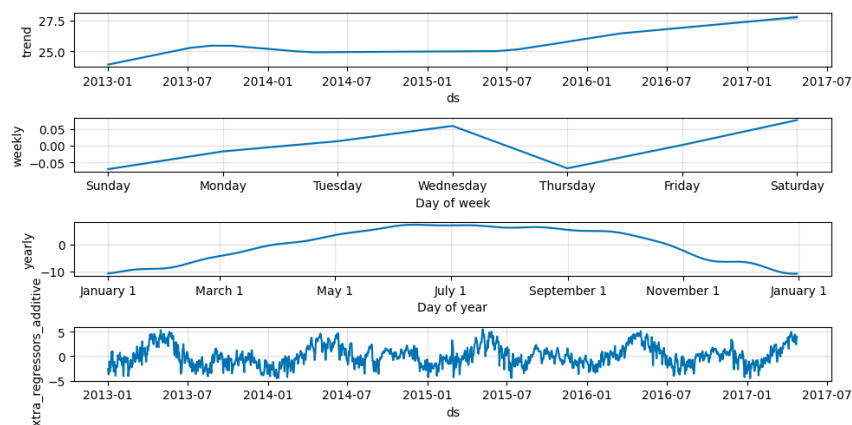
בהתבסס על מטריצת המתאם (Correlation Matrix) שהצגנו בחלק א, בחרנו שני המשתנים שהראו קשר הכי חזק עם הטמפרטורה, והשתמשנו ב "לחות יחסית" וב"מהירות הרוח" כ- משתנים מסבירים.

לצורך בניית מודל החיזוי, בחרנו ב-Prophet, אשר מתאים במיוחד לנתוני מזג האוויר שלנו, זאת משום שסדרות הזמן של המשתנים (כגון טמפרטורה, לחות ומהירות רוח) מציגות מגמות ארוכות טווח ועונתיות שנתית, ו-Prophet מתמודד עם מאפיינים אלה באמצעות פירוק רכיבי הסדרה.

בהתחלה הכנו את הנתונים, וידאנו שהמשתנה הנוסף נמצא במבנה הנתונים של סדרת הזמן שאין בו ערכים חסרים. אחר כך הוספנו את המשתנים למודל – קראנו לפונקציה " add regressor " כדי להגדיר שהמודל יקח בחשבון את המשתנים החיצוניים הללו.

אחרי זה אימון המודל, באמצעות קריאה לפונקציה "Fit", שבה הוא מזהה את הדפוסים על סמך הנתונים. לאחר אימון המודל, יצרנו סט נתונים חדש עם נקודות זמן עתידיות (כולל ערכים צפויים למשתנה החיצוני) והשתמשנו בפונקציה "Predict" כדי לקבל תחזיות.

אלו הם המרכיבים של התחזית שקיבלנו:



הגרף מציג את רכיבי המודל של Prophet לאחר שילוב המשתנים החיצוניים (Exogenous Variable), והוא מחולק לארבעה חלקים עיקריים: המגמה (Trend), העונתיות השנתית (Yearly Seasonality), העונתיות השבועית (Weekly Seasonality) והשפעת המשתנים החיצוניים (Extra Regressors Additive). כל אחד מהחלקים מסייע בהבנה מעמיקה יותר של דפוסי השינוי בטמפרטורה ומבהיר כיצד המשתנה החיצוני משפיע על התחזיות.

הגרף העליון מציג את מגמת הטמפרטורה לאורך השנים, ניתן להבחין כי הטמפרטורה נמצאת במגמת עלייה הדרגתית, מה שמעיד על מגמה ארוכת טווח של התחממות בדלהי.

הגרף האמצעי מתאר את התנודות החוזרות בטמפרטורה במהלך השנה, ניתן לראות שבחודשים הראשונים של השנה (ינואר-יולי) יש עלייה בטמפרטורה, ואילו מ-אוגוסט ואילך יש ירידה הדרגתית, התנהגות זו משקפת את עונות השנה האופייניות לדלהי: חורף קר, קיץ חם וגשום, ולאחר מכן התקררות בסוף השנה.

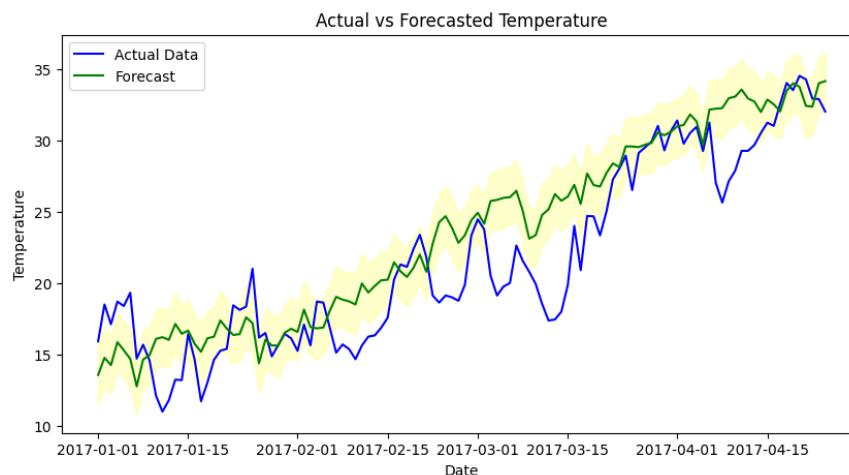
מחזוריות זו היא קריטית לתחזיות, שכן היא מאפשרת למודל לצפות תנודות צפויות בטמפרטורה בהתאם לעונות השנה.

הגרף התחתון מציג את התרומה של המשתנים החיצוניים לתחזית הטמפרטורה, ניתן לראות כי השפעת המשתנים משתנה בתדירות גבוהה לאורך התקופה, מה שמעיד על כך שהוא משחק תפקיד דינמי בתחזיות.

כאשר התנודות חזקות יותר, פירוש הדבר שהמשתנה החיצוני משפיע משמעותית על הטמפרטורה. לדוגמה, למשל ייתכן שבימים של לחות גבוהה יש השפעה חזקה יותר על תחושת החום והטמפרטורה בפועל.

כשלב אחרון, השוונו את הביצועים של המודל עם ה- Test, להערכת איכות החיזוי של המודל, השתמשנו בנתוני אמת של מזג האוויר מהתקופה 24/4/2017–1/1/2017 (Test), השוונו את תחזיות המודל לנתונים בפועל, הצגנו אותם על גבי אותו גרף, וחישבנו את סטייתם מהערכים האמיתיים.

הגרף למטה מציג את השוואת תחזיות הטמפרטורה של המודל (Prophet) לנתוני הטמפרטורה בפועל בין ינואר לאפריל 2017.



- ניתן לראות שהמודל מצליח ללכוד את מגמת העלייה הכללית של הטמפרטורה לאורך התקופה, ישנה התאמה טובה בין התחזיות לערכים בפועל, מה שמעיד על כך שהמודל קלט בצורה טובה את דפוסי השינוי בטמפרטורה.

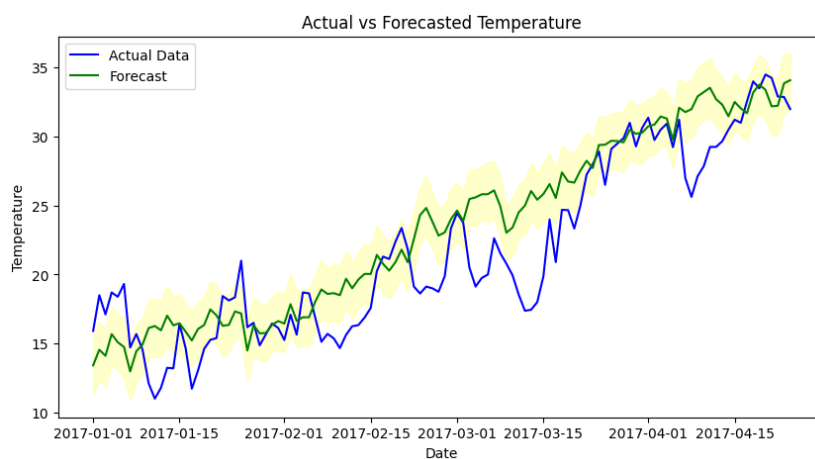
לבסוף, חישבנו מדדי שגיאה של התחזית מהנתונים האמיתיים:

ממוצע השגיאה - הערך המתקבל הוא -1.67, כלומר, באופן ממוצע, התחזיות של המודל נמוכות בכ- 1.67 מעלות מהערכים בפועל, המשמעות היא שהמודל מפגין הטיה קלה כלפי מטה, כלומר, יש לו נטייה להערכת חסר של הטמפרטורה.

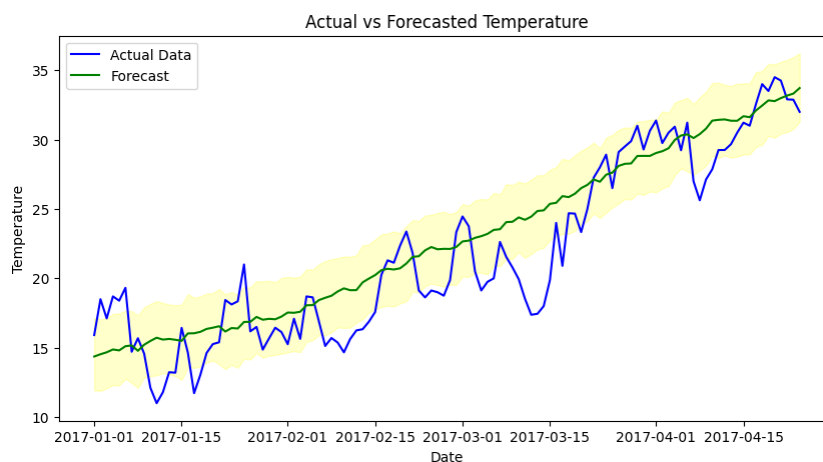
שונות השגיאה - הערך של 7.65 מצביע על פיזור די רחב של השגיאות סביב הממוצע, ערך גבוה יחסית של שונות עשוי להעיד על כך שהמודל מתקשה ללכוד שינויים חדים בטמפרטורה, מה שמתבטא בחיזויים עם רמות דיוק משתנות לאורך התקופה.

סיבה אפשרית לכך שהתוצאות החמירו ביחס למודל ללא המידע החיצונית היא העובדה שהשתמשנו ב-2 משתנים חיצוניים בעלי קורלציה הפוכה, אחד שלילי ואחד חיובי. בנוסף, לפי הגרף השלישי בניתוח תוצאות ה-PREDICT ראינו כי וקטור המשתנים מוסיף רעש בפני עצמו כאשר במחצית הראשונה של השנה הוא חיובי ובמחצית השנייה לרוב שלילי. החלטנו לבדוק בעקבות כך כל פרמטר בפני עצמו.

כאשר ביצענו חיזוי לטמפרטורה בהתבסס על משתנה הלחות בלבד ראינו כי התחזית הצליחה לעקוב אחרי המגמה הכללית של השינוי בטמפרטורה לאורך הזמן, דבר מעיד על זה שהלחות אכן משפיעה על הטמפרטורה ומהווה משתנה מסביר רלוונטי.



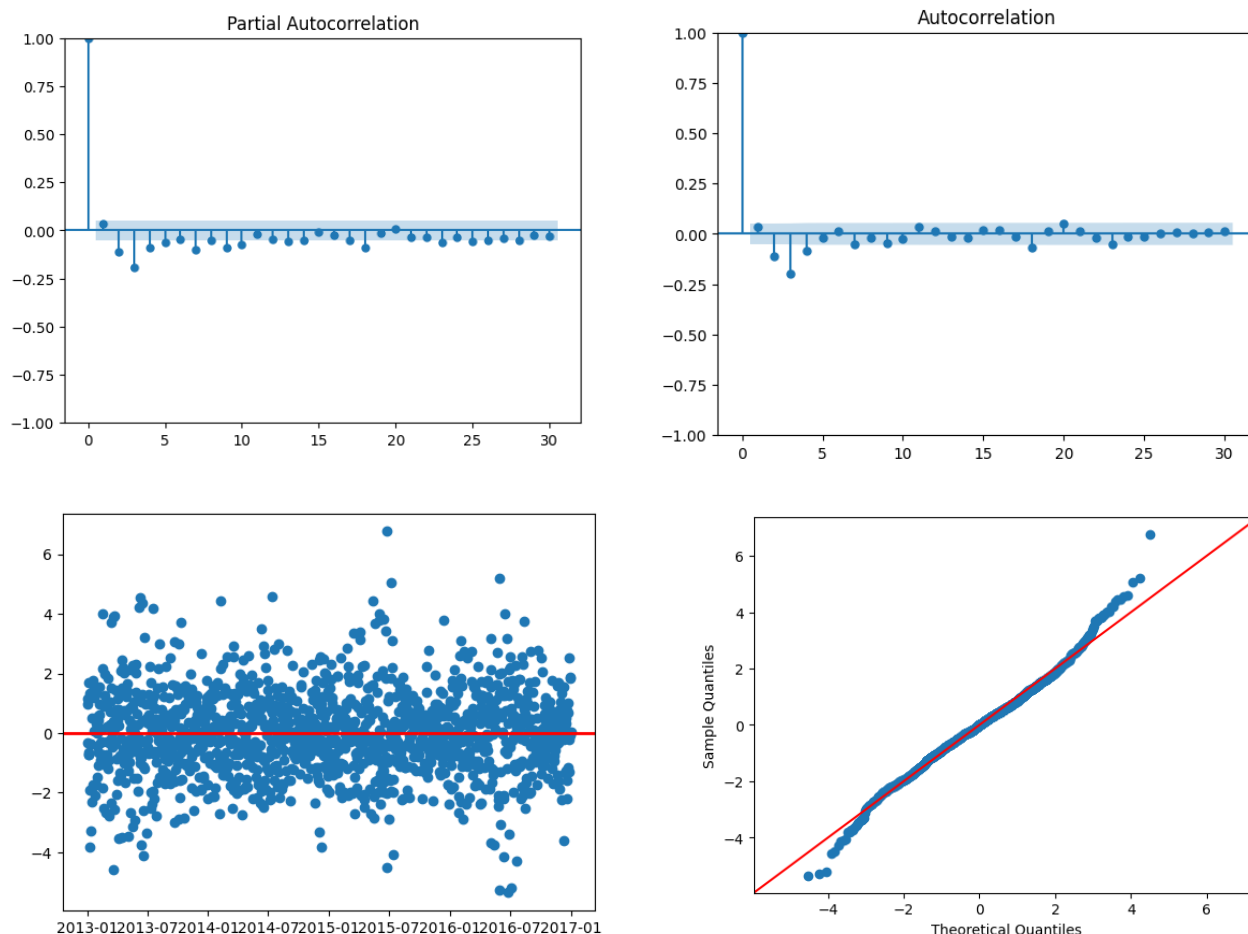
כאשר ביצענו חיזוי לטמפרטורה בהתבסס על משתנה מהירות הרוח בלבד, ניתן לראות כי התחזית אינה מצליחה ללכוד התנודות של הטמפרטורה בצורה טובה, הקו החזוי, מציג מגמה כללית עולה, אך הוא חלק מאוד ואינו משקף את השינויים החדים שנצפים בנתוני האמת.



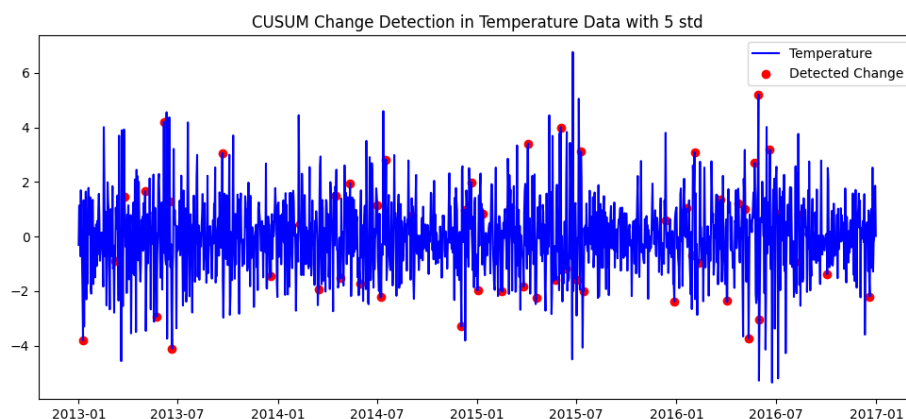
משני הגרפים אנחנו רואים כי המשתנה "לחות הממוצעת" הוא משתנה מסביר טוב יותר את השינויים בטמפרטורה מאשר המשתנה של "מהירות הרוח".

חלק 4

בחרנו להסתכל על השאריות של מודל Holt Winters כי הוא היה המוצלח ביותר. מבדיקה מקדימה אנחנו רואים כי הן נורמליות וסטציונריות לפי ACF, PACF, ומבחן ADF.



לפי ה-CUSUM, מתקבלות נקודות שינוי רבות עבור 3 ס"ת וגם עבור 5 ס"ת אם כי עבור 5 ס"ת משמעותית פחות. סיבות אפשריות לכך הן שונות גדולה במזג האוויר של דלהי והעובדה כי אנחנו משתמשים במידע יומי שמייצר רעש גדול יותר מאם היינו מסתכלות באופן שבועי.



אם נתבונן בנקודות השינוי עם 5 סטיות תקן נשים לב כי פרט לשנת 2015, ההפרש בין הנקודות הוא כ-3 חודשים, עם שינוי העונות. דבר זה יכול להיות הגיוני היות וכל עונה היא עם ממוצע טמפרטורה שונה ובנקודה מסוימת מזהים את השינוי בעונת השנה.

אחרי חיפוש התברר שבשנת 2015 היו אסונות טבע רבים מהרגיל ולפי אתרי חדשות מאותה שנה היא הייתה שנה קיצונית מבחינת חום [6],[7] מה שמסביר את הכמות הגדולה של נקודות השינוי שאנחנו רואים באותה מסגרת זה. באותם מקורות מסבירים כי באותה שנה היה השפעה חזרה של אל נינו, מחזור האקלימי של האוקיינוס הפסיפי, שהוביל לשינויים קיצוניים במזג האוויר. באותה שנה היו גם כמויות גדולות של הרוגים והרס בעקבות הקשיים [8],[9].

לסיכום, אנחנו רואים כי הטמפרטורה הממוצעת בדלהי משתנה רבות וקשה לזהות את התנהגותה ללא התבוננות כוללת של כל השנה. נראה כי השונות שלה באופן טבעי גדולה וכך גם של משתנים אחרים של מזג האוויר שמקשה על שילובם בשביל חיזוי. בנוסף ישנם משתנים נוספים משמעותיים מעבר למשתנים בנתונים שנאספו במסגרת הדאטה סט, כמו אל נינו, אשר משפיעים על מזג האוויר ובעלי חשיבות גדולה בהודו.

- [1] ABOUT DELHI - [Information about Delhi](#)
- [2] Climate and Average Weather Year Round in New Delhi - [New Delhi Climate, Weather By Month, Average Temperature \(NCT, India\) - Weather Spark](#)
- [3] NASA - [World of Change: Global Temperatures](#)
- [4] [20120105_PerceptionsAndDice.pdf](#)
- [5] [SARIMA model fit runs super slowly · Issue #6033 · statsmodels/statsmodels](#)
- [6] [2015 was India's third hottest year on record: IMD](#)
- [7] ['2015 was the warmest year on record in recent history' | Chandigarh News - The Indian Express](#)
- [8] [Gujarat floods: 72 people dead, over 81,000 cattle perished due to heavy rains – Firstpost](#)
- [9] [Northeast monsoon claimed 470 lives in Tamil Nadu: Jayalalithaa - The Hindu BusinessLine](#)