**Project Analysis Log**

**Objective**

The goal of this project is to predict **Average Stress and Emotional Overload** based on textual features, pregnancy-related variables, and sentiment analysis scores derived from Reddit posts.

**Initial Steps**

1. **Dataset Overview:**
   * The dataset consists of 100 rows with 27 columns, including:
     + Textual data (text column).
     + Sentiment scores (neg, neu, pos, compound).
     + Pregnancy-related variables such as Pregnancy Week (Model).
     + Stress intensity and emotional overload scores.
   * Target variable: **Average Stress and Emotional Overload**.
2. **Feature Engineering:**
   * TF-IDF vectorization of the text column to extract important words.
   * Additional features:
     + Text length (text\_length).
     + Number of question marks in the text (question\_marks).
     + Ratio of negative to positive sentiment (neg\_pos\_ratio).
3. **Modeling Approaches:**
   * Linear Regression.
   * Random Forest Regressor.
   * Recursive Feature Elimination (RFE) to select the top 10 features.

**Results and Observations**

**Model Evaluation**

1. **Linear Regression:**
   * Mean Squared Error (MSE): **6.79**.
   * R²: **-0.29** (negative R² indicates poor performance).
2. **Random Forest Regressor:**
   * Mean Squared Error (MSE): **8.54**.
   * R²: **-0.62** (indicating even worse performance compared to Linear Regression).

**Conclusion:** Both models failed to capture meaningful relationships between the features and the target variable. This suggests a need to improve feature selection, engineering, or try alternative models.

**Feature Importance (Random Forest):**

1. The top features selected by the model include words derived from TF-IDF:
   * from, since, first, on, up.
   * Importance scores indicate these features contribute the most to the model.
2. Pregnancy-related variables such as Pregnancy Week (Model) were **not selected as important** by RFE or Random Forest.

**Interpretation:**

* The lack of significance for Pregnancy Week (Model) suggests it may not have a strong linear or nonlinear relationship with the target variable. However, further investigation with engineered interactions or advanced models could provide more insight.

**Visualizations**

1. **Feature Importance Plot:**
   * Highlights the contribution of each feature, with TF-IDF words dominating.
2. **Predicted vs Actual Values (Linear Regression & Random Forest):**
   * Both plots show predictions far from the perfect prediction line, reflecting poor model performance.

**Next Steps**

**1. Try Advanced Models**

* Implement Gradient Boosting models such as:
  + XGBoost.
  + LightGBM.
* Investigate their ability to capture nonlinear relationships.

**2. Feature Engineering Improvements**

* Test interaction features:
  + Combining Pregnancy Week (Model) with sentiment scores.
  + Adding temporal features if timestamps are available.
* Perform additional text preprocessing:
  + Extract specific pregnancy-related keywords.
  + Use semantic embeddings (e.g., BERT) for richer text representation.

**3. Refine Target Variable**

* Analyze the target distribution and ensure it aligns with the prediction goals.
* Explore alternate targets such as binary stress classification (e.g., High vs Low).

**4. Reassess Pregnancy Week**

* Conduct focused analysis to determine:
  + Correlations between Pregnancy Week (Model) and the target.
  + Whether it becomes significant after adding interaction terms or nonlinear transformations.

**Challenges and Considerations**

1. Small dataset size (100 rows):
   * Limited data could restrict model performance.
   * Augmenting the dataset or using transfer learning approaches could help.
2. High-dimensional TF-IDF features:
   * Reducing dimensionality (e.g., PCA) or using domain-specific word embeddings might improve results.
3. Low R² scores:
   * Indicate the current features may not explain the target well, requiring deeper feature exploration.

**Conclusions**

* **Pregnancy Week (Model):**
  + Currently, it appears non-significant, but this does not rule out its potential if combined with other features or explored using nonlinear methods.
* **Next Actions:**
  + Focus on improving feature engineering.
  + Experiment with advanced models.
  + Document all observations to refine the process.

**Analysis Report for Gradient Boosting Models**

**Objective**

The goal of this stage was to improve the predictive performance of the model for estimating "Average Stress and Emotional Overload" by employing gradient boosting methods, specifically XGBoost and LightGBM.

**Process**

1. **Preprocessing and Feature Engineering**:
   * Dataset cleaned to remove rows with missing values in relevant columns.
   * Features added:
     + text\_length: Length of text posts.
     + question\_marks: Count of question marks as a proxy for emotional intensity.
     + neg\_pos\_ratio: Ratio of negative to positive sentiment.
   * TF-IDF applied to extract 100 most important textual features.
2. **Model Training**:
   * Split data into training (80%) and testing (20%) sets.
   * Models:
     + **XGBoost**: Optimized for regression tasks, trained with default hyperparameters.
     + **LightGBM**: A faster and efficient alternative, trained with default settings.
3. **Evaluation Metrics**:
   * Mean Squared Error (MSE): Measures error magnitude.
   * R2R^2R2: Explains the variance captured by the model.

**Results**

1. **XGBoost**:
   * MSE: 7.17
   * R2R^2R2: -0.36 (poor fit, indicates the model underperformed).
2. **LightGBM**:
   * MSE: 6.09
   * R2R^2R2: -0.16 (better than XGBoost but still indicates the model failed to generalize well).
3. **Feature Importances**:
   * LightGBM identified key features from the dataset, primarily from the TF-IDF text analysis.
   * Features like text\_length and question\_marks were moderately important.
4. **Visualizations**:
   * **Prediction vs Actuals**:
     + Both models struggled to align predictions with actual values, as shown in the scatter plots.
   * **Feature Importance Plot**:
     + LightGBM highlighted TF-IDF features as most significant.

**Conclusions**

1. **Model Performance**:
   * Both models underperformed, as indicated by negative R2R^2R2, suggesting the features do not sufficiently explain the variance in the target variable.
   * Despite LightGBM showing slightly better results, overall predictions deviate significantly from actual values.
2. **Feature Engineering**:
   * Features like Pregnancy Week (Model) and sentiment scores may not hold a strong linear or nonlinear relationship with the target variable in this setup.
   * TF-IDF features dominated the models, yet their explanatory power remains limited.
3. **Challenges**:
   * **Small dataset**: Only 100 samples were used, which limits the models' ability to generalize.
   * **High dimensionality**: TF-IDF features increased feature space significantly, potentially leading to overfitting.

**Next Steps**

1. **Model Improvement**:
   * Explore advanced neural architectures, such as BERT, for text representation.
   * Fine-tune hyperparameters for both XGBoost and LightGBM.
2. **Feature Engineering**:
   * Introduce interaction terms (e.g., combining Pregnancy Week with sentiment features).
   * Extract semantic embeddings using pre-trained language models.
3. **Dataset Augmentation**:
   * Increase the dataset size through additional scraping or synthetic data generation.
4. **Target Refinement**:
   * Simplify the prediction task by converting the target variable into binary classes (e.g., High vs Low stress).

This process highlights the complexity of modeling text-based predictions for stress and the necessity for richer features or larger datasets. Let me know if you'd like assistance implementing the next steps!

**דו"ח סיכום שלב ניתוח המודלים - Gradient Boosting**

**מטרה**

שיפור יכולת החיזוי של המודל עבור משתנה המטרה "רמת סטרס ממוצעת ועומס רגשי" באמצעות מודלים של Gradient Boosting - XGBoost ו-LightGBM.

**תהליך העבודה**

1. **עיבוד והנדסת תכונות**:
   * ניקוי הנתונים להסרת שורות עם ערכים חסרים בעמודות הרלוונטיות.
   * תכונות חדשות שנוספו:
     + text\_length: אורך הטקסט.
     + question\_marks: מספר סימני שאלה בטקסט (כמדד לעומס רגשי).
     + neg\_pos\_ratio: יחס בין סנטימנט שלילי לחיובי.
   * שימוש ב-TF-IDF כדי לנתח את הטקסט ולחלץ 100 תכונות משמעותיות.
2. **אימון המודלים**:
   * חלוקה לסט אימון (80%) וסט בדיקה (20%).
   * מודלים:
     + **XGBoost**: אומן כמודל רגרסיה עם פרמטרים בסיסיים.
     + **LightGBM**: מודל מהיר ויעיל יותר, אומן גם הוא עם פרמטרים בסיסיים.
3. **מדדי הערכה**:
   * MSE (שגיאה ממוצעת בריבוע): מודד את גודל השגיאה.
   * R2R^2R2: מודד את כמות השונות שהמודל מצליח להסביר.

**תוצאות**

1. **XGBoost**:
   * MSE: 7.17
   * R2R^2R2: -0.36 (התאמה נמוכה מאוד).
2. **LightGBM**:
   * MSE: 6.09
   * R2R^2R2: -0.16 (קצת טוב יותר מ-XGBoost, אך עדיין נמוך).
3. **חשיבות תכונות**:
   * LightGBM זיהה את התכונות המובילות, בעיקר מה-TF-IDF.
   * תכונות כמו text\_length ו-question\_marks קיבלו משקל בינוני.
4. **הדמיות**:
   * **גרף חיזוי לעומת ערכים אמיתיים**:
     + שני המודלים לא הצליחו לחזות בצורה מדויקת את הערכים האמיתיים.
   * **גרף חשיבות תכונות**:
     + תכונות TF-IDF בלטו כמשמעותיות ביותר.

**מסקנות**

1. **ביצועי המודלים**:
   * שני המודלים לא הצליחו לחזות את הערכים בצורה מדויקת, כפי שמשתקף מ-R2R^2R2 השלילי.
   * LightGBM הראה תוצאות מעט טובות יותר, אך עדיין הביצועים נמוכים.
2. **הנדסת תכונות**:
   * תכונות כמו Pregnancy Week (Model) וניקוד סנטימנט אינם מראים קשר חזק למשתנה המטרה.
   * תכונות ה-TF-IDF אומנם היו משמעותיות, אך תרומתן להסבר המשתנה המטרה מוגבלת.
3. **אתגרים**:
   * **גודל קטן של הדאטהסט**: 100 דוגמאות בלבד מקשות על המודלים ללמוד ולהכליל.
   * **מימדיות גבוהה**: כמות התכונות ש-TF-IDF הוסיפה עשויה לגרום ל-overfitting.

**שלבים הבאים**

1. **שיפור המודלים**:
   * לנסות שימוש במודלים מתקדמים כמו BERT לייצוג טקסט.
   * לבצע Fine-tuning למודלים כמו XGBoost ו-LightGBM.
2. **שיפור הנדסת תכונות**:
   * להוסיף אינטראקציות בין תכונות (לדוגמה, שילוב Pregnancy Week עם נתוני סנטימנט).
   * להעשיר את הטקסט על ידי Embeddings סמנטיים.
3. **הגדלת הדאטהסט**:
   * להגדיל את הדאטהסט על ידי איסוף מידע נוסף או יצירת נתונים סינתטיים.
4. **שינוי משתנה המטרה**:
   * להפוך את החיזוי לדו-ערכי (למשל: סטרס גבוה/נמוך) כדי לפשט את הבעיה.

שלב זה מדגיש את המורכבות בניתוח טקסט ואת הצורך בשיפור המודלים והתכונות. אני כאן אם תרצי עזרה בהמשך!

This stage evaluates the performance of the selected models (Linear Regression, Random Forest, and LightGBM) for predicting the target variable "Average Stress and Emotional Overload." Additionally, it assesses the models' predictions against actual values using test data and provides cross-validation metrics.

**Current Step Analysis**:

1. **Recursive Feature Elimination (RFE)**:
   * Identified the top 10 most predictive features: neg, pos, baby, having, her, into, no, now, time, and week.
2. **Train-Test Evaluation**:
   * Models were trained using the selected features and evaluated on the test set.
   * Results:
     + **Linear Regression**:
       - Mean Squared Error (MSE): **5.34**
       - R²: **-0.01** (indicating very poor fit on test data)
     + **Random Forest**:
       - MSE: **4.65**
       - R²: **0.12** (better than Linear Regression but still weak predictive power)
     + **LightGBM**:
       - MSE: **4.66**
       - R²: **0.12** (similar performance to Random Forest)
3. **Cross-Validation**:
   * Models were validated using 5-fold cross-validation to assess consistency:
     + **Linear Regression**:
       - Average MSE: **3.59**
       - Average R²: **0.12**
     + **Random Forest**:
       - Average MSE: **4.03**
       - Average R²: **0.00**
     + **LightGBM**:
       - Average MSE: **3.97**
       - Average R²: **0.01**
4. **Visualization Results**:
   * Scatter plots of predicted vs. actual values were generated:
     + **Linear Regression**:
       - Points show deviation from the perfect prediction line, indicating a lack of fit.
     + **LightGBM**:
       - Similar pattern to Linear Regression, but with slightly better alignment to the perfect prediction line.

**Conclusions**:

* LightGBM and Random Forest demonstrated weak but better predictive performance than Linear Regression. Both had R² values of 0.12 on the test set.
* Cross-validation highlighted inconsistency in model performance, with Linear Regression showing the best average R² (0.12).
* The predicted vs. actual scatter plots indicate significant deviations, suggesting the need for model tuning and/or further feature engineering.

**Hebrew Version**

**מטרה**:  
שלב זה מעריך את ביצועי המודלים (רגרסיה לינארית, יער אקראי ו-LightGBM) בניבוי המשתנה "ממוצע עומס רגשי ומתח". כמו כן, הוא בוחן את תחזיות המודלים מול הערכים בפועל ומשווה את המדדים על קבוצת המבחן ואימות צולב.

**ניתוח השלב הנוכחי**:

1. **בחירת מאפיינים (RFE)**:
   * זוהו 10 המאפיינים המשמעותיים ביותר לניבוי:  
     neg, pos, baby, having, her, into, no, now, time, week.
2. **הערכת מודלים על קבוצת מבחן**:
   * המודלים אומנו באמצעות המאפיינים שנבחרו והוערכו על קבוצת המבחן:
     + **רגרסיה לינארית**:
       - שגיאה ריבועית ממוצעת (MSE): **5.34**
       - R²: **-0.01** (מעיד על התאמה חלשה מאוד לנתוני המבחן)
     + **Random Forest**:
       - MSE: **4.65**
       - R²: **0.12** (התאמה טובה יותר מרגרסיה לינארית אך עדיין חלשה)
     + **LightGBM**:
       - MSE: **4.66**
       - R²: **0.12** (ביצועים דומים ל-Random Forest)
3. **אימות צולב**:
   * המודלים נבדקו באמצעות אימות צולב של 5 קפלים:
     + **רגרסיה לינארית**:
       - MSE ממוצע: **3.59**
       - R² ממוצע: **0.12**
     + **Random Forest**:
       - MSE ממוצע: **4.03**
       - R² ממוצע: **0.00**
     + **LightGBM**:
       - MSE ממוצע: **3.97**
       - R² ממוצע: **0.01**
4. **תוצאות ויזואליזציה**:
   * גראפים של תחזיות לעומת ערכים בפועל:
     + **רגרסיה לינארית**:
       - הנקודות מציגות סטייה משמעותית מקו התחזית המושלם, מה שמצביע על חוסר התאמה.
     + **LightGBM**:
       - דפוס דומה לרגרסיה לינארית, עם יישור טוב מעט יותר לקו התחזית המושלם.

**מסקנות**:

* LightGBM ו-Random Forest הפגינו ביצועים חלשים אך טובים יותר מרגרסיה לינארית, עם R² של 0.12 על קבוצת המבחן.
* אימות צולב הדגיש חוסר עקביות בביצועי המודלים, כאשר רגרסיה לינארית הציגה את R² הממוצע הטוב ביותר (0.12).
* הגרפים של תחזיות לעומת ערכים בפועל מצביעים על סטיות משמעותיות, מה שמרמז על הצורך בכוונון פרמטרים ו/או שיפורי הנדסת מאפיינים.

רשת ניורונים

**עברית**

1. **תיאור המודל**:
   * המודל משתמש ברשת עצבית רבת שכבות לניתוח נתונים של חוויות הקשורות להריון.
   * הנתונים עובדו מראש ונוספו להם מאפיינים חשובים כגון יחס חיובי-שלילי, אורך טקסט, ואינטראקציות בין משתני ההריון והרגש השלילי.
2. **תוצאות אימון המודל**:
   * גרף עקומת אובדן (Loss Curve) מראה ירידה יציבה בערכי האובדן, הן בסט האימון והן בסט הוולידציה, אך נראה כי ישנן תנודות מסוימות המעידות על צורך באיזון בין אימון לבין מניעת אובראפיטינג.
   * לאחר 50 איטרציות, ערכי האובדן הסופי (MSE) על סט הבדיקה עמדו על 8.38, עם מדד R² של -0.59.
3. **ניתוח תוצאות**:
   * השוואת ערכים אמיתיים מול ערכים מנובאים (גרף Predictions vs Actuals) מדגימה פיזור מסוים שמרמז על דיוק נמוך במודל.
   * הערכים המנובאים קרובים לקו האדום רק במעט מקרים, מה שמעיד על כך שיש מקום לשיפור.
4. **ניתוח SHAP**:
   * ניתוח השפעות המאפיינים (SHAP) נתקל בשגיאה: הייתה אי-התאמה בין ממדי המאפיינים לבין הערכים של SHAP. יש לוודא את יישור הנתונים בין הסטים.
5. **המלצות לשיפור**:
   * לשפר את תהליך היישור בין המאפיינים לערכי SHAP.
   * לשקול להוסיף שכבת רגולריזציה כדי למנוע אובראפיטינג.
   * לבחון את השפעת הפיצ'רים על ידי הורדת משתנים מיותרים.

**English**

1. **Model Description**:
   * The model employs a multi-layer neural network to analyze pregnancy-related experiences.
   * The data underwent preprocessing with feature engineering, including positive-negative ratios, text length, and interactions between pregnancy weeks and negative sentiment.
2. **Model Training Results**:
   * The loss curve demonstrates a steady decline in loss values for both training and validation sets, with slight oscillations, suggesting potential overfitting.
   * After 50 epochs, the test set's final MSE was 8.38, with an R² score of -0.59.
3. **Result Analysis**:
   * The Predictions vs. Actuals plot indicates some degree of scattering, revealing low predictive accuracy.
   * Predicted values align with the red line only in a few cases, indicating room for improvement.
4. **SHAP Analysis**:
   * SHAP analysis encountered an alignment error: there was a mismatch between the feature dimensions and SHAP values. Data alignment between sets must be reviewed.
5. **Recommendations for Improvement**:
   * Ensure proper alignment between features and SHAP values to avoid errors.
   * Introduce a regularization layer to prevent overfitting.
   * Perform feature selection to remove irrelevant variables and improve model accuracy.

**מסקנות עד עכשיו**

**איזה מודל הכי טוב?**

1. **ביצועי הרשת העצבית**:
   * המודל המבוסס על רשת עצבית אינו מדויק במיוחד, כפי שמצביעים ערכי ה-MSE וה-R²:
     + MSE: 8.38
     + R²: -0.59 (ערך שלילי מעיד על כך שהמודל פחות טוב אפילו ממודל בסיסי שמחזיר ממוצע בלבד).
   * הגרף של "Predictions vs Actuals" מראה פיזור גדול, כלומר חוסר יכולת לנבא בצורה מדויקת את הערכים.
2. **השוואה למודלים אחרים**:
   * אין מידע מלא על מודלים אחרים שהופעלו בפרויקט. אם הוחלו מודלים פשוטים יותר כמו רגרסיה לינארית או מודלים מבוססי החלטות (Random Forest, XGBoost), ייתכן שהם יכולים להראות ביצועים טובים יותר.
   * יתרון של מודלים מבוססי רשתות עצביות לרוב מתקבל כשיש כמות נתונים גדולה מאוד, מה שלא נראה המקרה כאן.

**האם שבוע ההריון משפיע?**

1. **ניתוח ראשוני**:
   * בוצע הנדסת מאפיינים שכללה אינטראקציה בין שבוע ההריון לרגשות שליליים (week\_neg\_interaction), שהיו חלק מהמודל.
   * ניתוח SHAP, שתכליתו למדוד את השפעת הפיצ'רים, נכשל עקב בעיה ביישור נתונים. לכן, כרגע אין תוצאה ברורה לגבי השפעת שבוע ההריון.
2. **פעולות להמשך**:
   * לתקן את בעיית היישור של SHAP ולהריץ שוב את הניתוח.
   * לבחון השפעה ישירה של שבוע ההריון באמצעות מודלים פשוטים כמו רגרסיה לינארית עם מאפיין זה בלבד.

**מסקנות כלליות**

* **דיוק המודל**: הרשת העצבית אינה מספקת דיוק גבוה. נדרשת בחינה מחודשת של איכות הפיצ'רים.
* **השפעת שבוע ההריון**: עד כה, אין הוכחה חד-משמעית שהשפעת שבוע ההריון משמעותית, אך יש לבצע בדיקות נוספות.
* **בחירת מודל**: ייתכן שמודלים פשוטים יותר כמו רגרסיה או Random Forest יתנו תוצאות טובות יותר על סט הנתונים הנוכחי.

**Steps Moving Forward**

1. **Feature Evaluation**:
   * Check if features like "Pregnancy Week" or engineered features like "week\_neg\_interaction" improve prediction significantly.
   * Perform feature selection or ranking to identify the most influential variables.
2. **Model Optimization**:
   * Try simpler models such as linear regression or decision trees to compare performance.
   * Optimize the current neural network with techniques like hyperparameter tuning or regularization layers.
3. **Analysis of Pregnancy Week Impact**:
   * Run a dedicated analysis (e.g., regression) with "Pregnancy Week" as a primary predictor to directly measure its influence.
4. **SHAP Correction**:
   * Fix the data alignment error in SHAP analysis to better interpret feature contributions.

**השוואה למודלים אחרים: ניתוח ביצועים של רשת נוירונים מול מודלים קלאסיים ומתקדמים**

**תיאור המודלים והשוואה**

1. **רשת נוירונים (Neural Network):**
   * **מבנה:** רשת עצבית מרובת שכבות עם שכבות הסתר ותהליכי הנדסת תכונות (כגון TF-IDF, יחס סנטימנט שלילי-חיובי, ואינטראקציות).
   * **ביצועים:**
     + MSE (שגיאה ממוצעת בריבוע): **8.38**
     + R²: **-0.59**
   * **חוזקות:**
     + גרף עקומת האובדן (Loss Curve) מצביע על ירידה יציבה באובדן במהלך האימון.
     + גמישות בלמידת קשרים מורכבים בין מאפיינים.
   * **חולשות:**
     + ביצועים נמוכים יחסית על קבוצת המבחן.
     + פיזור גבוה בגרף Predictions vs Actuals.
     + בעיות יישור נתונים בערכי SHAP.
2. **מודלים קלאסיים:**
   * **רגרסיה לינארית:**
     + MSE: **5.34**
     + R²: **-0.01**
     + **חולשות:** מודל קו לינארי פשוט שאינו מצליח לתפוס קשרים מורכבים בין תכונות למשתנה המטרה.
   * **Random Forest:**
     + MSE: **4.65**
     + R²: **0.12**
     + **חוזקות:** גמישות בלמידת קשרים לא-לינאריים, אך מושפע ממגבלות על ממדיות גבוהה.
   * **LightGBM:**
     + MSE: **4.66**
     + R²: **0.12**
     + **חוזקות:** מהיר ויעיל לעבודה עם נתונים גבוהי ממדיות, עם ביצועים מעט טובים יותר מ-Random Forest.

**השוואת תובנות מרכזיות**

* **דיוק התחזיות:**
  + רשת הנוירונים הציגה ביצועים נמוכים משמעותית (R² שלילי), בעוד ש-LightGBM ו-Random Forest הציגו R² חיובי קטן.
  + Linear Regression נכשלה באופן ברור, עם R² שלילי שמעיד על אי-התאמה למידע.
* **חשיבות מאפיינים:**
  + ב-LightGBM ו-Random Forest, תכונות מבוססות TF-IDF שלטו בחשיבותן.
  + רשת הנוירונים, אף שלא יישרה את נתוני SHAP בהצלחה, עשויה להציע תובנות מעמיקות יותר לאחר תיקון בעיות האופטימיזציה.
* **הבדלים במורכבות המודלים:**
  + רשת הנוירונים מסוגלת ללמוד קשרים מורכבים יותר, אך דורשת כמות נתונים משמעותית גדולה יותר, בניגוד למודלים הקלאסיים שהתמודדו טוב יותר עם גודל מדגם קטן.

**מסקנות**

* **LightGBM ו-Random Forest:**  
  מציעים ביצועים טובים יחסית במודלים קלאסיים, אך עדיין אינם מצליחים לספק דיוק משמעותי.
* **רשת נוירונים:**  
  למרות הביצועים הנמוכים הנוכחיים, היא מראה פוטנציאל גבוה יותר ללמידת קשרים מורכבים אם תטופלנה הבעיות הבאות:
  + גודל מדגם קטן.
  + חוסר יישור נתוני SHAP.
  + תהליך רגולריזציה למניעת אובראפיטינג.

**המלצות לעתיד**

1. **שיפור הנדסת תכונות:**
   * יצירת אינטראקציות מותאמות בין "שבוע ההריון" למאפיינים רגשיים.
   * שימוש במודלים מתקדמים לניתוח טקסט, כמו BERT, להפקת מידע סמנטי.
2. **הגדלת הדאטהסט:**
   * איסוף נתונים נוספים כדי לשפר את למידת הקשרים.
3. **שימוש בשיטות מתקדמות:**
   * Fine-tuning של מודלים כמו LightGBM.
   * בדיקת רשתות עצביות מעמיקות כמו GPT או BERT להפקת תובנות איכותיות יותר.

**סיכום**

* נכון להיום, LightGBM מספקת את הפתרון הטוב ביותר, אך הפוטנציאל של רשת הנוירונים גבוה יותר בטווח הארוך.
* שבוע ההריון אינו מהווה משתנה משפיע משמעותי במודלים הנוכחיים, אך ייתכן שיש לו תרומה בקשרים מורכבים שעדיין לא זוהו.

**בדיקת קשר בין שבוע ההריון לבין הלחץ ללא מודל**  
  
הסתכלנו על הנתונים שהיו לנו ובדקנו בכלל האם יש קשר לאחר שראינו שמודלים לא עובדים

**מתאם פירסון**: ערך נמוך של -0.09, מה שמצביע על קשר חלש מאוד ולא משמעותי בין שבוע ההריון למדד הלחץ.

1. **מתאם ספירמן**: ערך של -0.12 עם p-value = 0.2305, כך שאין מובהקות סטטיסטית. המשמעות היא שהסדר היחסי של הנתונים אינו מצביע על קשר משמעותי.
2. **מתאם קנדל**: ערך של -0.09 עם p-value = 0.1962, מה שמחזק את המסקנה שאין קשר מובהק סטטיסטית בין שבוע ההריון למדד הלחץ.

**מסקנות:**

אין עדות סטטיסטית ברורה לקשר בין שבוע ההריון למדד הלחץ או העומס הרגשי, על פי הנתונים שבידיך.

 **פירוש הגרפים**:

* **גרף הפיזור (Scatter Plot)**: נראה שאין דפוס ברור המצביע על קשר לינארי בין שבוע ההריון למדד הסטרס. Pregnancy\_Week\_vs\_Stress\_Scatter
* **Box Plot לפי חלוקת שבועות**: המדדים מפוזרים די שווה בין קבוצות השבועות השונות, ואין מגמה ברורה של עלייה או ירידה ברמת הסטרס. Stress\_Distribution\_Across\_Pregnancy\_Weeks

בחינת משתנים אחרים והקשר שלהם ללחץ בשביל שיעזרו בחיזוי

לאחר ששבוע ההריון לא עזר רצינו להסתכל על משתנים אחרים ובאמת לראות מה יעזור לנו בחיזוי.

**נתונים כלליים**

הניתוח מבוסס על מדגם שכלל את העמודות הרלוונטיות: טקסט, שבוע ההיריון, ציוני רגשות (שליליים, ניטרליים, חיוביים, ומדד כולל), ומדד הלחץ הממוצע. בנוסף, בוצעו הנדסת מאפיינים שהוסיפה עמודות כגון: אורך הטקסט, מספר סימני שאלה, יחס רגשות שליליים-חיוביים, ואינטראקציה בין שבוע ההיריון לציונים השליליים.

**גרפים וניתוחים מרכזיים:**

1. **קשר בין שבוע ההיריון לבין הלחץ הממוצע**
   * מתאם פירסון: -0.09
   * מתאם ספירמן: -0.12 (לא מובהק סטטיסטית)
   * מתאם קנדל: -0.09 (לא מובהק סטטיסטית)

אין עדות לקשר חזק או מובהק סטטיסטית בין שבוע ההיריון לבין הלחץ הממוצע.

1. **רגשות שליליים (neg)**
   * מתאם פירסון: 0.34
   * מתאם ספירמן: 0.44 (מובהק סטטיסטית, P < 0.05)
   * מתאם קנדל: 0.32 (מובהק סטטיסטית, P < 0.05)

יש קשר חיובי ומובהק בין רגשות שליליים ללחץ הממוצע.

1. **רגשות ניטרליים (neu)**
   * מתאם פירסון: -0.24
   * מתאם ספירמן: -0.39 (מובהק סטטיסטית, P < 0.05)
   * מתאם קנדל: -0.28 (מובהק סטטיסטית, P < 0.05)

יש קשר שלילי ומובהק בין רגשות ניטרליים ללחץ הממוצע.

1. **רגשות חיוביים (pos)**
   * מתאם פירסון: -0.08
   * מתאם ספירמן: -0.03 (לא מובהק סטטיסטית)
   * מתאם קנדל: -0.02 (לא מובהק סטטיסטית)

אין עדות לקשר חזק או מובהק בין רגשות חיוביים ללחץ הממוצע.

1. **מדד כולל (compound)**
   * מתאם פירסון: -0.26
   * מתאם ספירמן: -0.29 (מובהק סטטיסטית, P < 0.05)
   * מתאם קנדל: -0.21 (מובהק סטטיסטית, P < 0.05)

יש קשר שלילי ומובהק בין המדד הכולל ללחץ הממוצע.

1. **אורך הטקסט (text\_length)**
   * מתאם פירסון: 0.31
   * מתאם ספירמן: 0.35 (מובהק סטטיסטית, P < 0.05)
   * מתאם קנדל: 0.25 (מובהק סטטיסטית, P < 0.05)

יש קשר חיובי ומובהק בין אורך הטקסט ללחץ הממוצע.

1. **מספר סימני שאלה (question\_marks)**
   * מתאם פירסון: -0.14
   * מתאם ספירמן: -0.25 (מובהק סטטיסטית, P < 0.05)
   * מתאם קנדל: -0.19 (מובהק סטטיסטית, P < 0.05)

יש קשר שלילי ומובהק בין מספר סימני שאלה ללחץ הממוצע.

1. **יחס רגשות שליליים-חיוביים (neg\_pos\_ratio)**
   * מתאם פירסון: -0.19
   * מתאם ספירמן: 0.30 (מובהק סטטיסטית, P < 0.05)
   * מתאם קנדל: 0.22 (מובהק סטטיסטית, P < 0.05)

יש קשר מובהק (אך לא חזק) בין היחס ללחץ הממוצע.

1. **אינטראקציה בין שבוע ההיריון לרגשות שליליים (week\_neg\_interaction)**
   * מתאם פירסון: 0.12
   * מתאם ספירמן: 0.16 (לא מובהק סטטיסטית)
   * מתאם קנדל: 0.11 (לא מובהק סטטיסטית)

אין עדות לקשר מובהק בין אינטראקציה זו ללחץ הממוצע.

**סיכום**

* הרגשות השליליים, הניטרליים, ואורך הטקסט הראו קשרים חזקים ומובהקים סטטיסטית עם מדד הלחץ הממוצע.
* שבוע ההיריון עצמו לא הציג קשר מובהק ללחץ.
* ממצאים אלו עשויים להוביל לפיתוח מודלים מדויקים יותר בעזרת שילוב מאפיינים רגשיים וטקסטואליים.

לכל אחד גם עשינו גרף מתאים של catters

כעת אנחנו יודעים איזה פיצ׳רים לכלול  
כדי לחזות את רמת הלחץ הממוצעת, כדאי להתמקד במאפיינים שהראו קשרים חזקים ומובהקים סטטיסטית עם מדד הלחץ. הנה סיכום המאפיינים שכדאי להסתמך עליהם:

**מאפיינים מרכזיים לחיזוי רמת הלחץ:**

1. **רגשות שליליים (neg)**
   * קשר חזק ומובהק (מתאמי ספירמן וקנדל מובהקים סטטיסטית).
   * רגשות שליליים משקפים בצורה ישירה את רמת הלחץ.
2. **רגשות ניטרליים (neu)**
   * קשר שלילי ומובהק.
   * ייתכן שהיעדר תוכן רגשי מעיד על חוסר התמודדות ישירה עם הלחץ.
3. **מדד כולל (compound)**
   * קשר שלילי ומובהק.
   * מסכם את כל הציונים הרגשיים, ולכן מהווה מדד מקיף וחשוב.
4. **אורך הטקסט (text\_length)**
   * קשר חיובי ומובהק.
   * טקסטים ארוכים עשויים להעיד על התבטאות מפורטת יותר של לחץ.
5. **מספר סימני שאלה (question\_marks)**
   * קשר שלילי ומובהק.
   * שאלות עשויות להעיד על חוסר וודאות או חיפוש מידע, מה שמשפיע על רמת הלחץ.
6. **יחס רגשות שליליים-חיוביים (neg\_pos\_ratio)**
   * קשר מובהק (ספירמן וקנדל).
   * יחס זה מצביע על מידת הדומיננטיות של רגשות שליליים ביחס לחיוביים, ומהווה מדד חשוב.

**מאפיינים שכדאי לשלב אך עם פחות דגש:**

1. **אינטראקציה בין שבוע ההיריון לרגשות שליליים (week\_neg\_interaction)**
   * לא נמצא קשר מובהק סטטיסטית, אך ייתכן שיש בו חשיבות כחלק ממודל רב-משתני.
2. **שבוע ההיריון עצמו (Pregnancy Week (Model))**
   * לא נמצא קשר מובהק ישיר, אך ייתכן שהוא ישפיע כאשר הוא משולב עם מאפיינים אחרים.

**אסטרטגיה לחיזוי:**

1. **מודל רגרסיה או מודלים מבוססי למידת מכונה (Machine Learning)**:
   * מודלים כמו Gradient Boosting, Random Forest, או Neural Networks עשויים להתאים במיוחד.
   * כלים כמו XGBoost או LightGBM יכולים לעבוד היטב עם מאפיינים מספריים וטקסטואליים.
2. **אימות וייעול המודל**:
   * חשוב לבדוק את ביצועי המודל באמצעות חלוקה לסט אימון וסט בדיקה.
   * מדדים כמו Mean Absolute Error (MAE) או R² יכולים לשמש להערכת ביצועי המודל.