**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.