Work Productivity & Influencing Factors Analysis

PRESENTATION - HIEN HOANG

25 MAY - 2025

TABLE OF CONTENT

Project Objectives

- Understand the factors impacting actual employee productivity.
- Build a predictive model to estimate individual productivity scores based on behavioral and perception data.
- Deliver actionable insights to support HR and business strategies for workforce development.

Data Overview

Source: https://www.kaggle.com/datasets/mahdimashayekhi/social-media-vs-productivity/data

Dataset size

30,000 records (simulated/real employee responses)

Variables analyzed

- Behavioral: Social media usage, sleep hours, stress level, work hours, use of focus apps, etc.
- Perceptual: Perceived productivity, job satisfaction.
- Target: Actual productivity score (0–10 scale)

Goal

Predict and explain what drives productivity differences among employees.

```
The dataset has 30000 rows and 19 columns.
Dataset columns:
 Index(['age', 'gender', 'job_type', 'daily_social_media_time',
       'social_platform_preference', 'number_of_notifications',
       'work_hours_per_day', 'perceived_productivity_score',
       'actual_productivity_score', 'stress_level', 'sleep_hours',
       'screen_time_before_sleep', 'breaks_during_work', 'uses_focus_apps',
       'has_digital_wellbeing_enabled', 'coffee_consumption_per_day',
       'days_feeling_burnout_per_month', 'weekly_offline_hours',
       'job_satisfaction_score'],
      dtype='object')
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 30000 entries, 0 to 29999
Data columns (total 19 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   age                             30000 non-null  int64
 1   gender                          30000 non-null  object
 2   job_type                        30000 non-null  object
 3   daily_social_media_time         27235 non-null  float64
 4   social_platform_preference      30000 non-null  object
 5   number_of_notifications         30000 non-null  int64
 6   work_hours_per_day              30000 non-null  float64
 7   perceived_productivity_score    28386 non-null  float64
 8   actual_productivity_score       27635 non-null  float64
 9   stress_level                    28096 non-null  float64
 10  sleep_hours                     27402 non-null  float64
 11  screen_time_before_sleep        27789 non-null  float64
 12  breaks_during_work              30000 non-null  int64
 13  uses_focus_apps                 30000 non-null  bool
 14  has_digital_wellbeing_enabled   30000 non-null  bool
 15  coffee_consumption_per_day      30000 non-null  int64
 16  days_feeling_burnout_per_month  30000 non-null  int64
 17  weekly_offline_hours            30000 non-null  float64
 18  job_satisfaction_score          27270 non-null  float64
dtypes: bool(2), float64(9), int64(5), object(3)
```

```python
# Kiểm tra giá trị thiếu trong từng cột
missing_values = df.isnull().sum()
print("Missing values per column:\n", missing_values)

# Xử lý các cột số (numeric columns)
# Khởi tạo SimpleImputer với chiến lược mean (thay thế bằng mean)
numeric_columns = df.select_dtypes(include=['float64']).columns
imputer_numeric = SimpleImputer(strategy='mean')

# Áp dụng SimpleImputer vào các cột số
df[numeric_columns] = imputer_numeric.fit_transform(df[numeric_columns])

# Kiểm tra lại giá trị thiếu sau khi xử lý
missing_values_after = df.isnull().sum()
print("\nMissing values after imputation:\n", missing_values_after)
```

```
Missing values per column:
 age                                0
gender                              0
job_type                            0
daily_social_media_time          2765
social_platform_preference          0
number_of_notifications             0
work_hours_per_day                  0
perceived_productivity_score     1614
actual_productivity_score        2365
stress_level                     1904
sleep_hours                      2598
screen_time_before_sleep         2211
breaks_during_work                  0
uses_focus_apps                     0
has_digital_wellbeing_enabled       0
coffee_consumption_per_day          0
days_feeling_burnout_per_month      0
weekly_offline_hours                0
job_satisfaction_score           2730
dtype: int64
```

Analytical Process

```python
# Các cột số cần kiểm tra outlier
numeric_columns = [
    'daily_social_media_time', 'weekly_offline_hours', 'work_hours_per_day',
    'perceived_productivity_score', 'actual_productivity_score',
    'stress_level', 'sleep_hours', 'screen_time_before_sleep',
    'job_satisfaction_score'
]

def check_outliers_iqr(data, col):
    Q1 = data[col].quantile(0.25)
    Q3 = data[col].quantile(0.75)
    IQR = Q3 - Q1
    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR
    outliers = data[(data[col] < lower_bound) | (data[col] > upper_bound)]
    print(f"Column: {col}")
    print(f"  Lower bound: {lower_bound:.3f}")
    print(f"  Upper bound: {upper_bound:.3f}")
    print(f"  Number of outliers: {outliers.shape[0]}")
    print(f"  Percentage of outliers: {100 * outliers.shape[0] / data.shape[0]:.2f}%\n")

# Kiểm tra outlier cho từng cột
for col in numeric_columns:
    check_outliers_iqr(df, col)
```

```
Column: daily_social_media_time
  Lower bound: -1.835
  Upper bound: 7.852
  Number of outliers: 348
  Percentage of outliers: 1.16%

Column: weekly_offline_hours
  Lower bound: -11.597
  Upper bound: 31.439
  Number of outliers: 116
  Percentage of outliers: 0.39%

Column: work_hours_per_day
  Lower bound: 1.577
  Upper bound: 12.421
  Number of outliers: 97
  Percentage of outliers: 0.32%

Column: perceived_productivity_score
  Lower bound: -1.106
  Upper bound: 12.132
  Number of outliers: 0
  Percentage of outliers: 0.00%

Column: actual_productivity_score
  Lower bound: -0.821
  Upper bound: 10.727
  Number of outliers: 0
  Percentage of outliers: 0.00%

Column: stress_level
  Lower bound: -4.500
  Upper bound: 15.500
  Number of outliers: 0
  Percentage of outliers: 0.00%

Column: sleep_hours
  Lower bound: 2.899
  Upper bound: 10.099
  Number of outliers: 0
  Percentage of outliers: 0.00%

Column: screen_time_before_sleep
  Lower bound: -0.730
  Upper bound: 2.735
  Number of outliers: 198
  Percentage of outliers: 0.66%

Column: job_satisfaction_score
  Lower bound: -0.789
  Upper bound: 10.725
  Number of outliers: 0
  Percentage of outliers: 0.00%
```

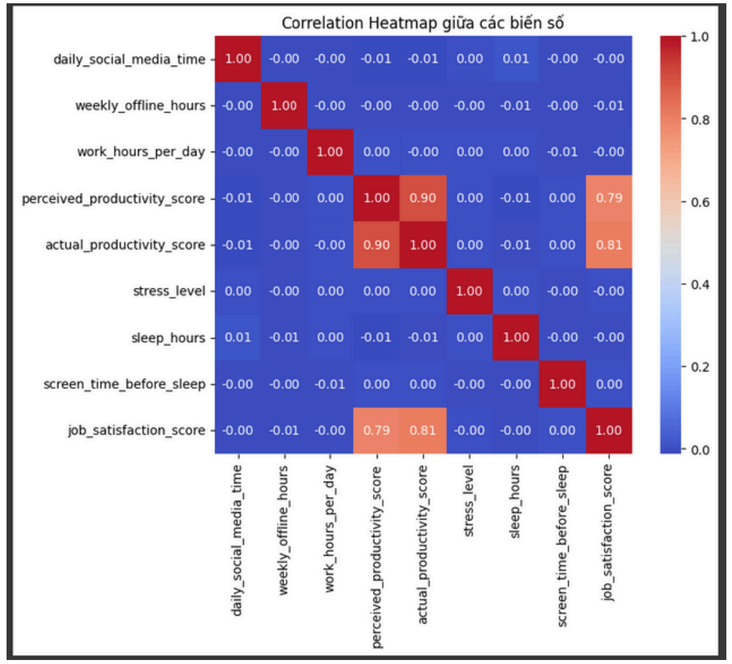
```python
# Kiểm tra số lượng bản ghi trùng lặp
num_duplicates = df.duplicated().sum()
print(f"Số bản ghi trùng lặp: {num_duplicates}")

# Loại bỏ các bản ghi trùng lặp
df_no_duplicates = df.drop_duplicates()
print(f"Số bản ghi sau khi loại bỏ trùng lặp: {df_no_duplicates.shape[0]}")
```

```
Số bản ghi trùng lặp: 0
Số bản ghi sau khi loại bỏ trùng lặp: 30000
```

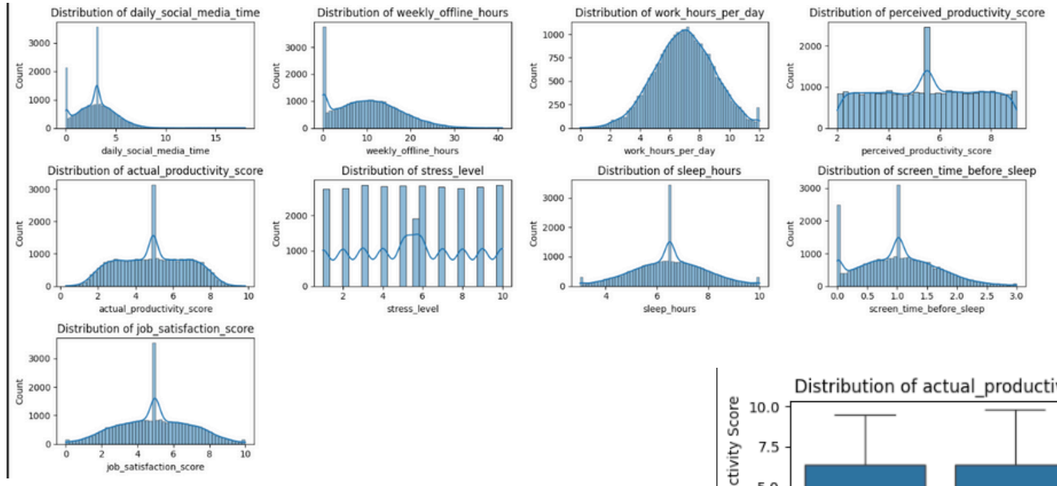
Data Cleaning & Preparation

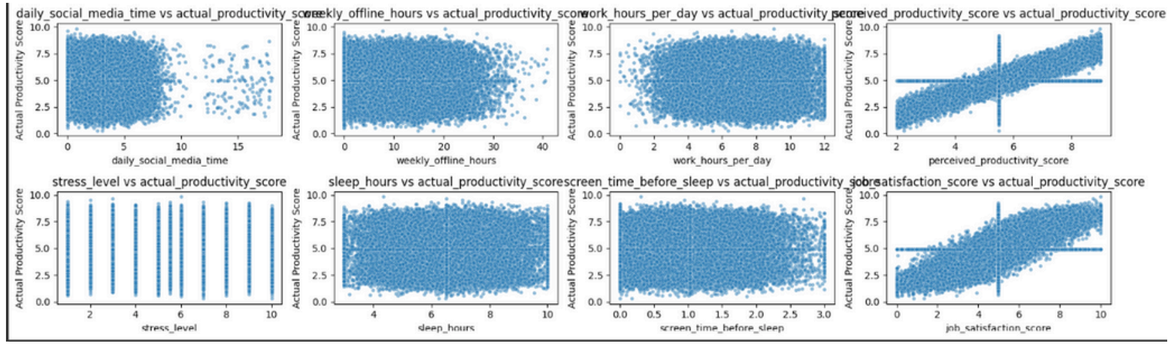
- Handled missing values
- Ensured correct data types
- Check duplicate
- Outliers (Numberic comlumns)

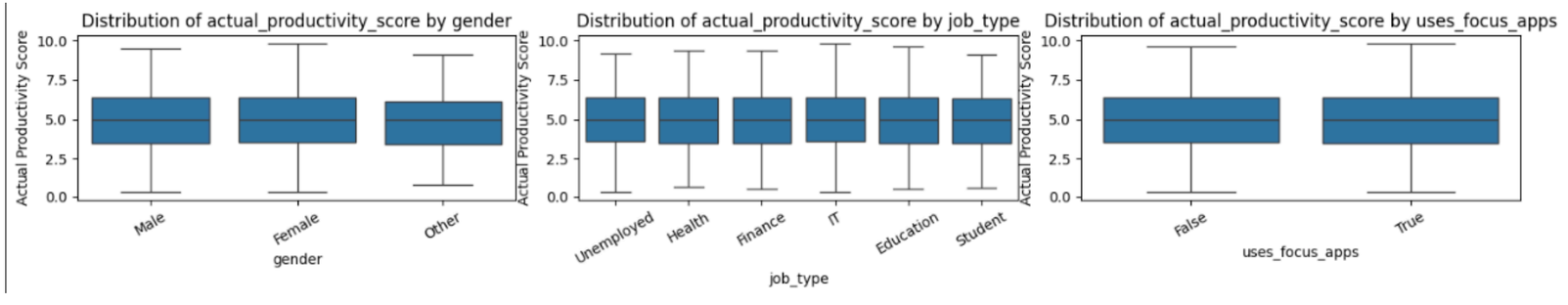


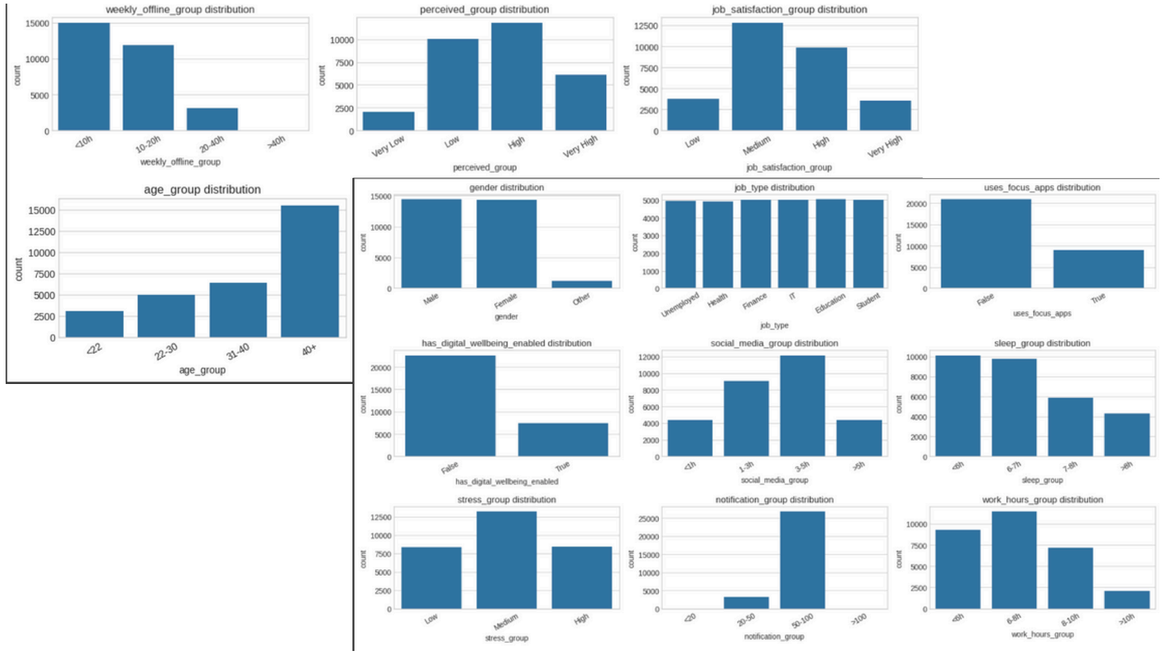
Analytical Process

Exploratory Data Analysis of numerical features (EDA 1)



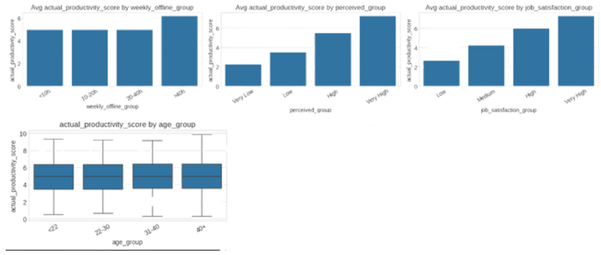


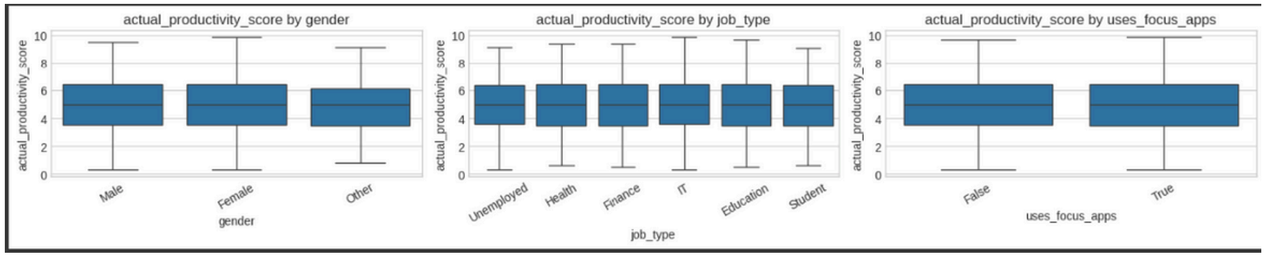


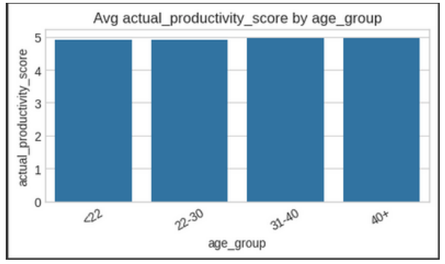


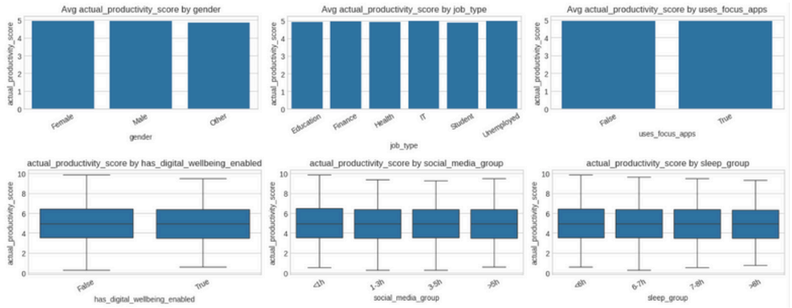
Analytical Process

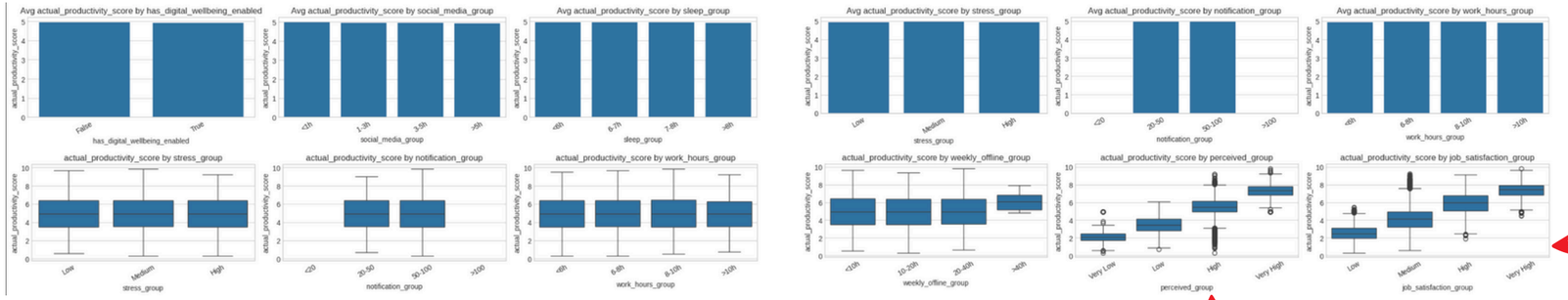
Exploratory Data Analysis of categorical features (EDA 2)











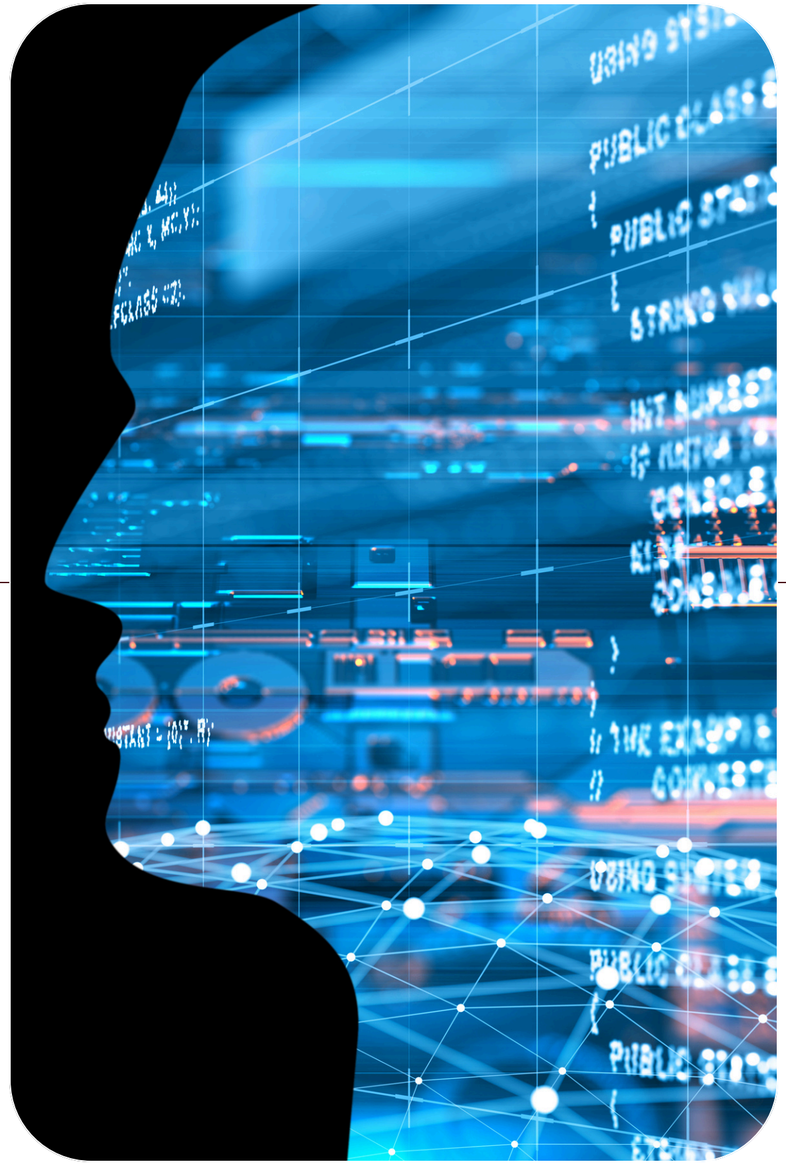
EDA – Overall conclusion

- Perceived productivity and job satisfaction are the strongest determinants of actual productivity in this dataset.
- Objective behavioral factors have little to no significant impact.
- There are no high-risk or extreme productivity groups in the data; the dataset is quite homogeneous with few outliers.
- Internal programs should focus on improving engagement, job satisfaction, and productivity awareness rather than simply controlling surface-level behaviors.

1. Numerical Variables
   - Most numerical variables have a near-normal or slightly right-skewed distribution, with few outliers or extreme values.
   - “Perceived productivity score” and “job satisfaction score” show a very strong correlation with “actual productivity score”.
   - This suggests that personal perception of performance and job satisfaction are the strongest predictors of actual productivity.
   - Behavioral variables such as social media usage, sleep hours, stress level, and work hours per day show weak or no clear relationship with actual productivity.
   - No other numerical variable besides perceived productivity and job satisfaction stands out in distinguishing productivity scores.

2. Categorical Variables
   - Category distributions are generally balanced, with a very small percentage of extreme behavior groups (e.g., >5h social media, >8h sleep, >100 notifications).
   - Actual productivity score shows minimal differences across gender, job type, and focus app usage groups.
   - Only perceived group and job satisfaction group show clear separation in productivity scores.
   - Behavioral grouping (social, sleep, stress, work hours, etc.) does not result in significant differences in productivity scores.

Predictive Modeling Approach



1. Feature Selection
- Selected four main predictors based on EDA and domain knowledge:
- Perceived productivity score
- Job satisfaction score
- Sleep hours
- Daily social media time

2. Data Preparation
- Defined feature matrix (X) and target variable (y: actual productivity score).
- Split data into train (80%) and test (20%) sets to evaluate generalizability.

3. Model Building
- Model used: Random Forest Regression (n_estimators=100, random_state=42).
- Trained the model on the training set and evaluated predictions on the test set.

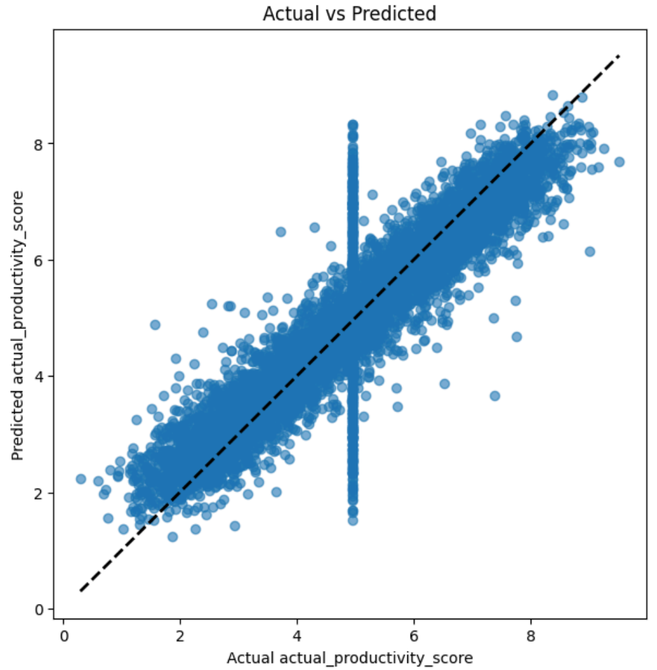
4. Model Evaluation
- Metrics:
- Root Mean Squared Error (RMSE)
- Mean Absolute Error (MAE)
- R-squared (R^2)
- Result:
- The model achieves high accuracy, with low error and high R^2.

5. Feature Importance Analysis
- Assessed feature importance from the trained Random Forest model.
- Perceived productivity and job satisfaction are the strongest predictors of actual productivity.

6. Model Optimization
- Performed hyperparameter tuning using GridSearchCV (n_estimators, max_depth).
- Selected the best model based on lowest RMSE and highest R^2.

7. Visualization
- Visualized feature importances (horizontal bar plot).
- Plotted Actual vs. Predicted scores to assess model fit.

Model Performance & Feature Importance



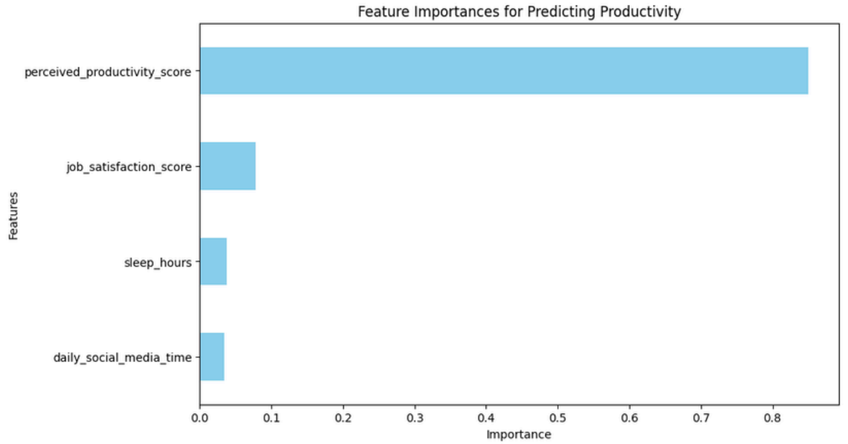
Baseline Model Performance

Metric	Value
RMSE	0.72
MAE	0.52
R^2	0.84

- The Random Forest model achieves high predictive accuracy with an R^2 score of 0.84, indicating that 84% of the variance in productivity scores is explained by the model.
- The MAE of 0.52 and RMSE of 0.72 demonstrate that prediction errors are small on average, confirming the model's practical utility.

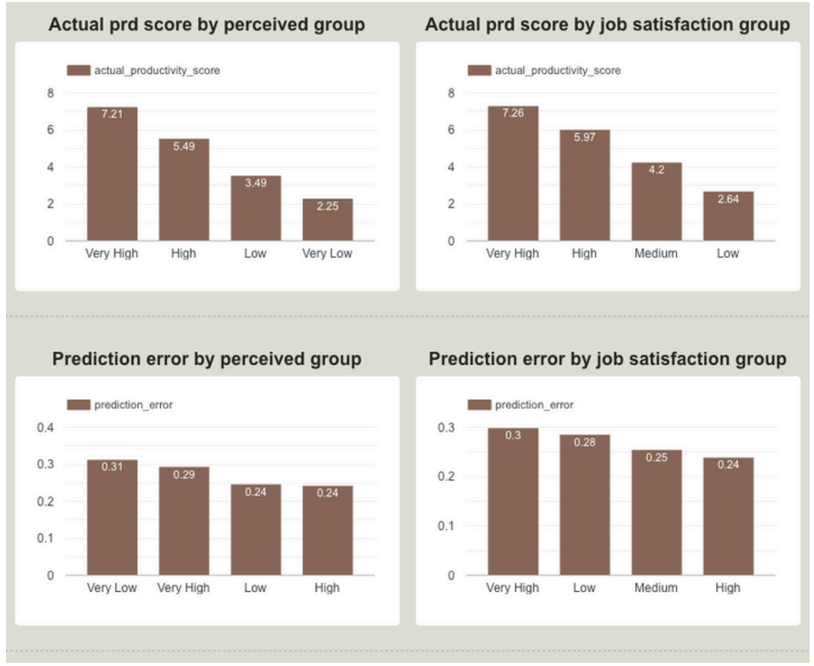
- Optimized Model Performance (GridSearchCV)
  - Best Parameters: max_depth=10, n_estimators=200
  - Optimized RMSE: 0.70
  - Optimized R^2: 0.85

After hyperparameter tuning using GridSearchCV, the optimized Random Forest model achieved improved performance with an RMSE of 0.70 and R^2 of 0.85. The best parameters found were max_depth=10 and n_estimators=200. This indicates that model tuning contributed to slightly more accurate predictions of productivity scores.

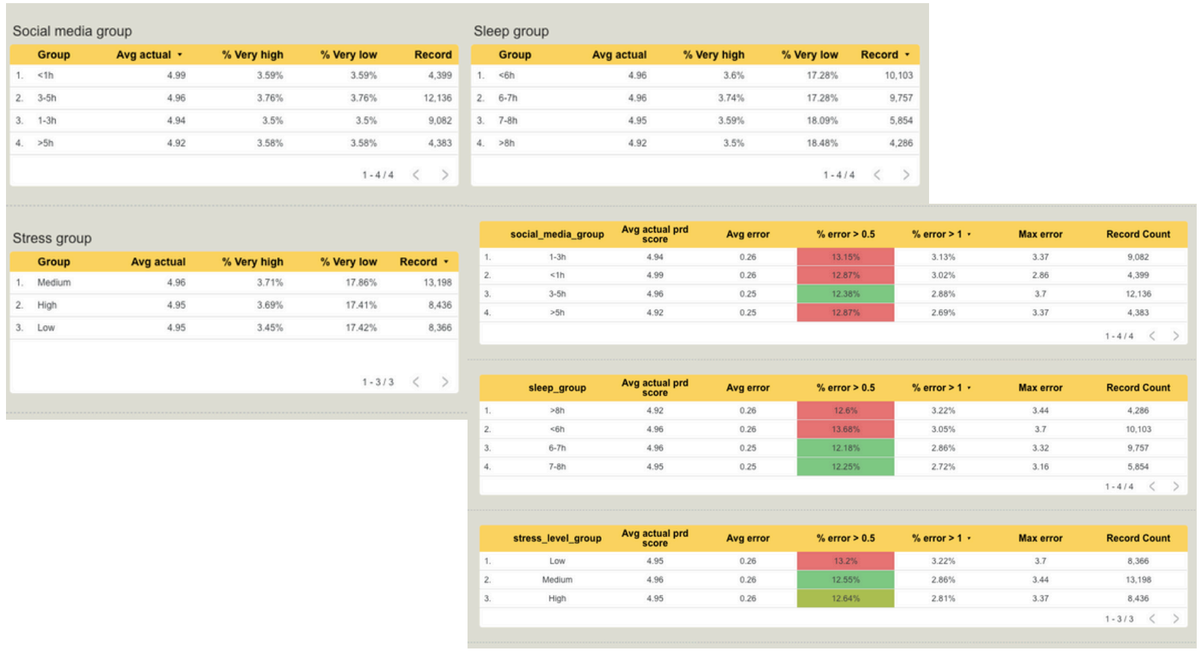


What Drives Productivity?

Visualization & Key Insights



- Key Drivers of Productivity:
  - Self-reported factors (perceived productivity, job satisfaction) have the strongest impact on actual productivity.
  - Productivity sharply increases with higher perceived and job satisfaction groups.
  - Prediction errors are lowest for highly satisfied groups.
- Behavioral Group Analysis:
  - Social media usage, sleep, stress: No group stands out as having significantly higher or lower productivity.
  - Distribution of "very high" or "very low" productivity is similar across behavioral groups.
- Model Robustness:
  - Prediction errors are consistently low across all behavioral groups.
  - No specific group produces extreme outliers in model performance.



Social media group

	Group	Avg actual	% Very high	% Very low	Record
1.	<1h	4.99	3.59%	3.59%	4,399
2.	3-5h	4.96	3.76%	3.76%	12,136
3.	1-3h	4.94	3.5%	3.5%	9,082
4.	>5h	4.92	3.58%	3.58%	4,383

1 - 4 / 4

Sleep group

	Group	Avg actual	% Very high	% Very low	Record
1.	<6h	4.96	3.6%	17.28%	10,103
2.	6-7h	4.96	3.74%	17.28%	9,757
3.	7-8h	4.95	3.59%	18.09%	5,854
4.	>8h	4.92	3.5%	18.48%	4,286

1 - 4 / 4

Stress group

	Group	Avg actual	% Very high	% Very low	Record
1.	Medium	4.96	3.71%	17.86%	13,198
2.	High	4.95	3.69%	17.41%	8,436
3.	Low	4.95	3.45%	17.42%	8,366

1 - 3 / 3

	social_media_group	Avg actual prd score	Avg error	% error > 0.5	% error > 1	Max error	Record Count
1.	1-3h	4.94	0.26	13.15%	3.13%	3.37	9,082
2.	<1h	4.99	0.26	12.87%	3.02%	2.86	4,399
3.	3-5h	4.96	0.25	12.38%	2.88%	3.7	12,136
4.	>5h	4.92	0.25	12.87%	2.69%	3.37	4,383

1 - 4 / 4

	sleep_group	Avg actual prd score	Avg error	% error > 0.5	% error > 1	Max error	Record Count
1.	>8h	4.92	0.26	12.6%	3.22%	3.44	4,286
2.	<6h	4.96	0.26	13.68%	3.05%	3.7	10,103
3.	6-7h	4.96	0.25	12.18%	2.86%	3.32	9,757
4.	7-8h	4.95	0.25	12.25%	2.72%	3.16	5,854

1 - 4 / 4

	stress_level_group	Avg actual prd score	Avg error	% error > 0.5	% error > 1	Max error	Record Count
1.	Low	4.95	0.26	13.2%	3.22%	3.7	8,366
2.	Medium	4.96	0.26	12.55%	2.86%	3.44	13,198
3.	High	4.95	0.26	12.64%	2.81%	3.37	8,436

1 - 3 / 3

Top 3 Practical Applications for Businesses

1/ Focus on Employee **Engagement & Satisfaction**

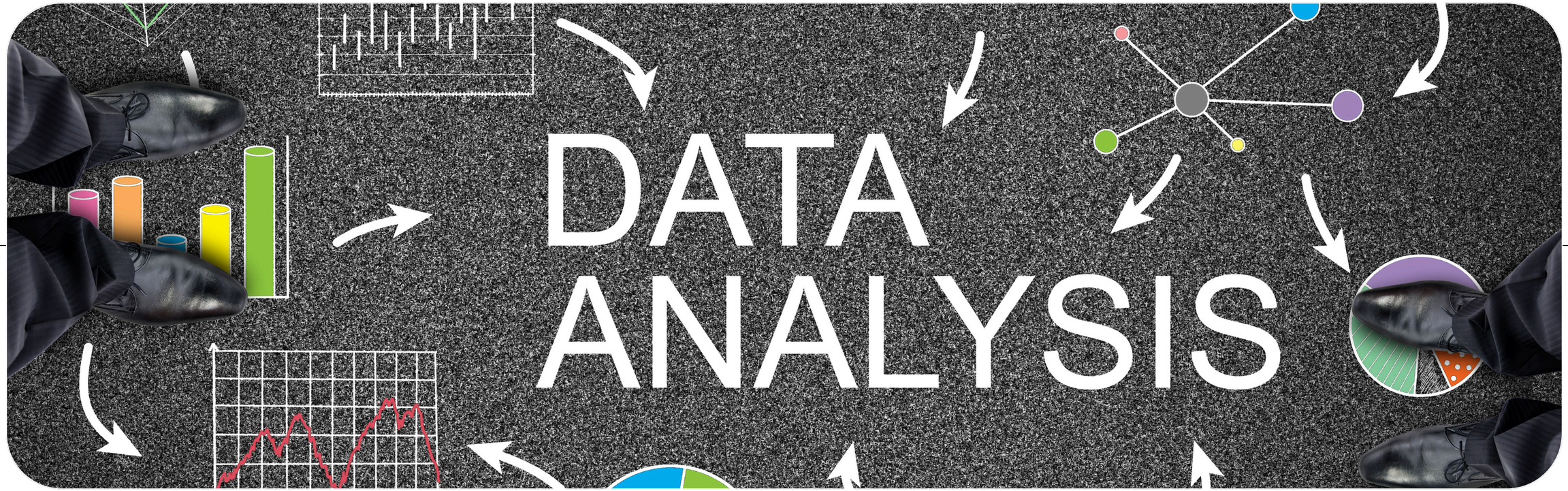
- Directly invest in initiatives that boost job satisfaction and perceived productivity =>Actions: Employee recognition programs, transparent communication, regular feedback, career development planning.

2/ Implement Predictive Monitoring Systems

- Use the predictive model to identify individuals or teams at risk of low productivity early => Actions: Integrate **survey/feedback** data with HR analytics platforms; set up alerts for declining engagement or satisfaction scores.

3/ Design Targeted HR Interventions

- Develop **personalized support plans** (e.g., coaching, flexible work arrangements, tailored wellness programs) for groups identified as vulnerable => Actions: Data-driven workshops, stress management resources, mentoring for low-engagement employees.



Limitations — Key Points

1. Heavy Reliance on **Self-Reported Data**
    - Model results may be affected by subjective perceptions, which do not always accurately reflect actual productivity.
2. **Limited and Cross-Sectional** Dataset
    - Only a few predictors and a single point in time were analyzed, limiting our ability to capture all factors and observe long-term trends.
3. **Limited Generalizability**
    - Findings may not be applicable to all industries, regions, or employee types due to sample and feature constraints.

Q & A

THANK YOU

Unlock the Power of Data!

HIEN HOANG