CS 4372.501 - Assignment 1

Anurag Nagar - Fall 2025

Garments Worker Productivity Prediction

Student: Vaishnavi Pasumarthi, Akshay Nagarajan

NetID: VXP210093, AHN210000

Due Date: September 21, 2025

Dataset and Methodology

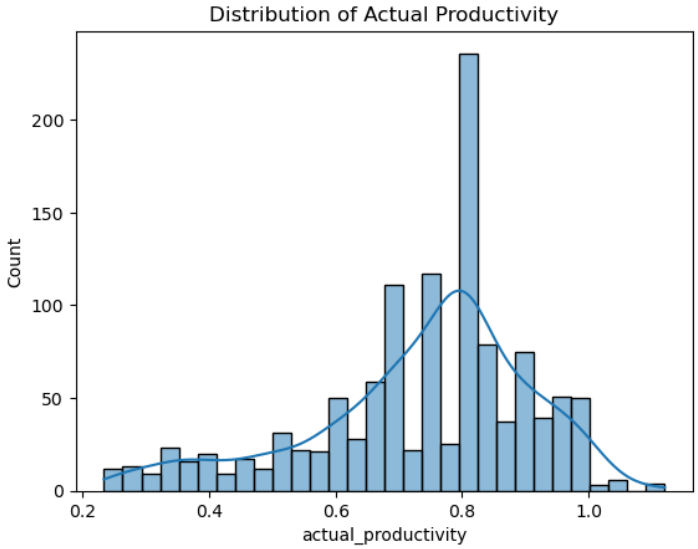
This analysis uses the Garments Worker Productivity dataset from the UCI Machine Learning Repository. The target variable actual_productivity represents worker performance in garment manufacturing. The Standard Deviation for this feature is 0.174488.

Data preprocessing included handling missing values, extracting date-based features (month and weekday), creating ratio and per-worker features, standardizing numeric features, generating polynomial interaction terms for selected variables, and one-hot encoding categorical variables, all combined into a unified pipeline for train/test data preparation.

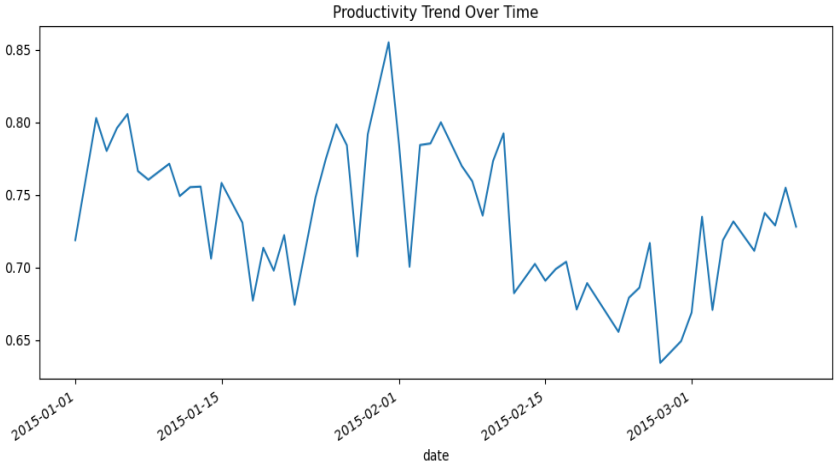
Attribute Analysis

The dataset contained 1,197 records with 15 features, of which 4 were categorical (date, quarter, department, day) and 11 were numerical.

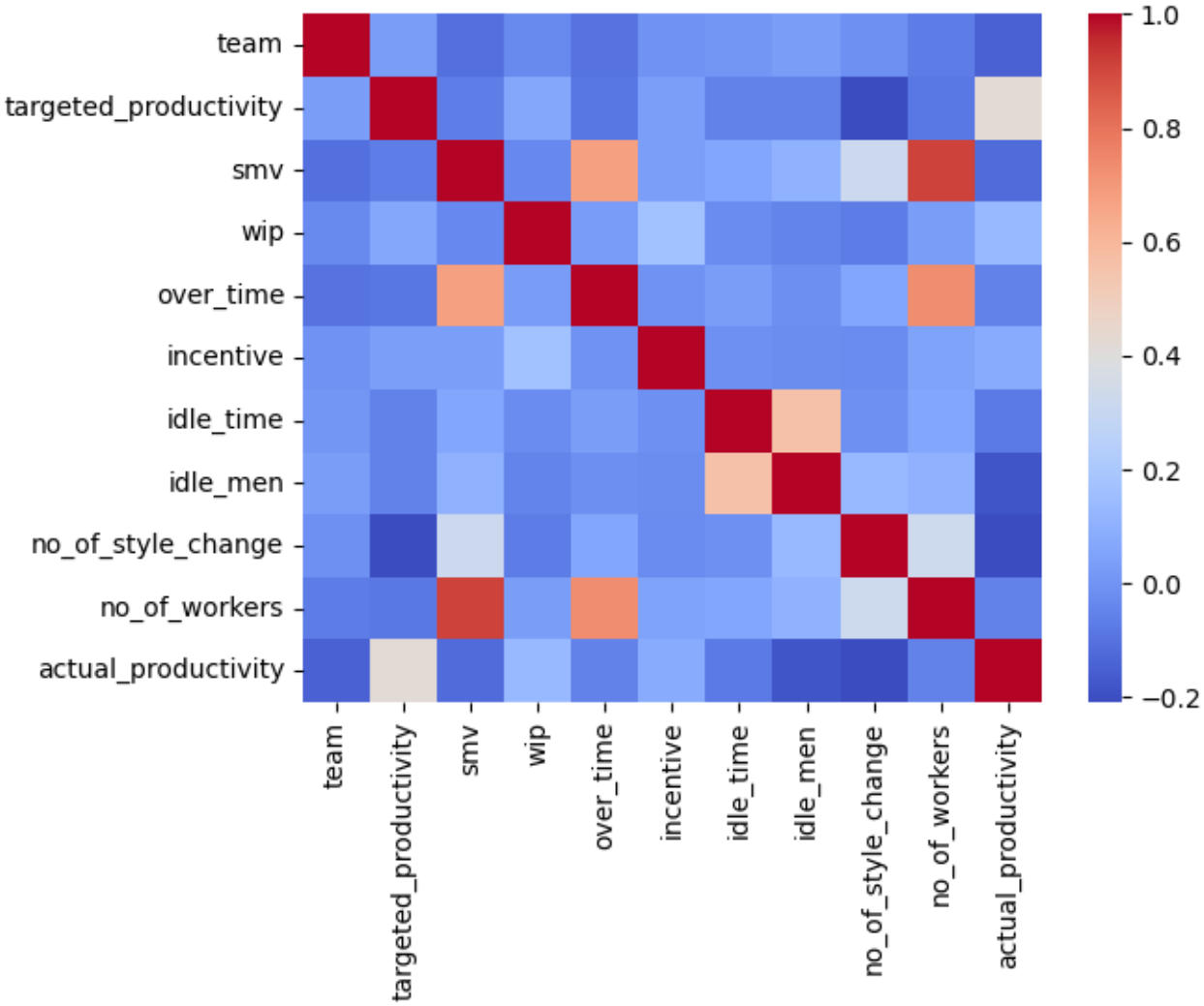
- Missing Values: The column wip had 506 missing entries (42.27%), requiring special treatment during preprocessing.
- Target Variable – actual_productivity:
  - Count: 1,197
  - Mean: 0.735
  - Standard Deviation: 0.174
  - Minimum: 0.234, Maximum: 1.120
  - Quartiles: 25% = 0.650, 50% = 0.773, 75% = 0.850
  - Skewness: -0.81 (left-skewed, most values are high)
  - Kurtosis: 0.33 (slightly flatter than normal distribution)



- Trend Analysis: Productivity values generally clustered between 0.65 and 0.85, showing stable performance around the target. Dips below 0.5 indicated underperformance, while outliers above 1.0 reflected overachievement.

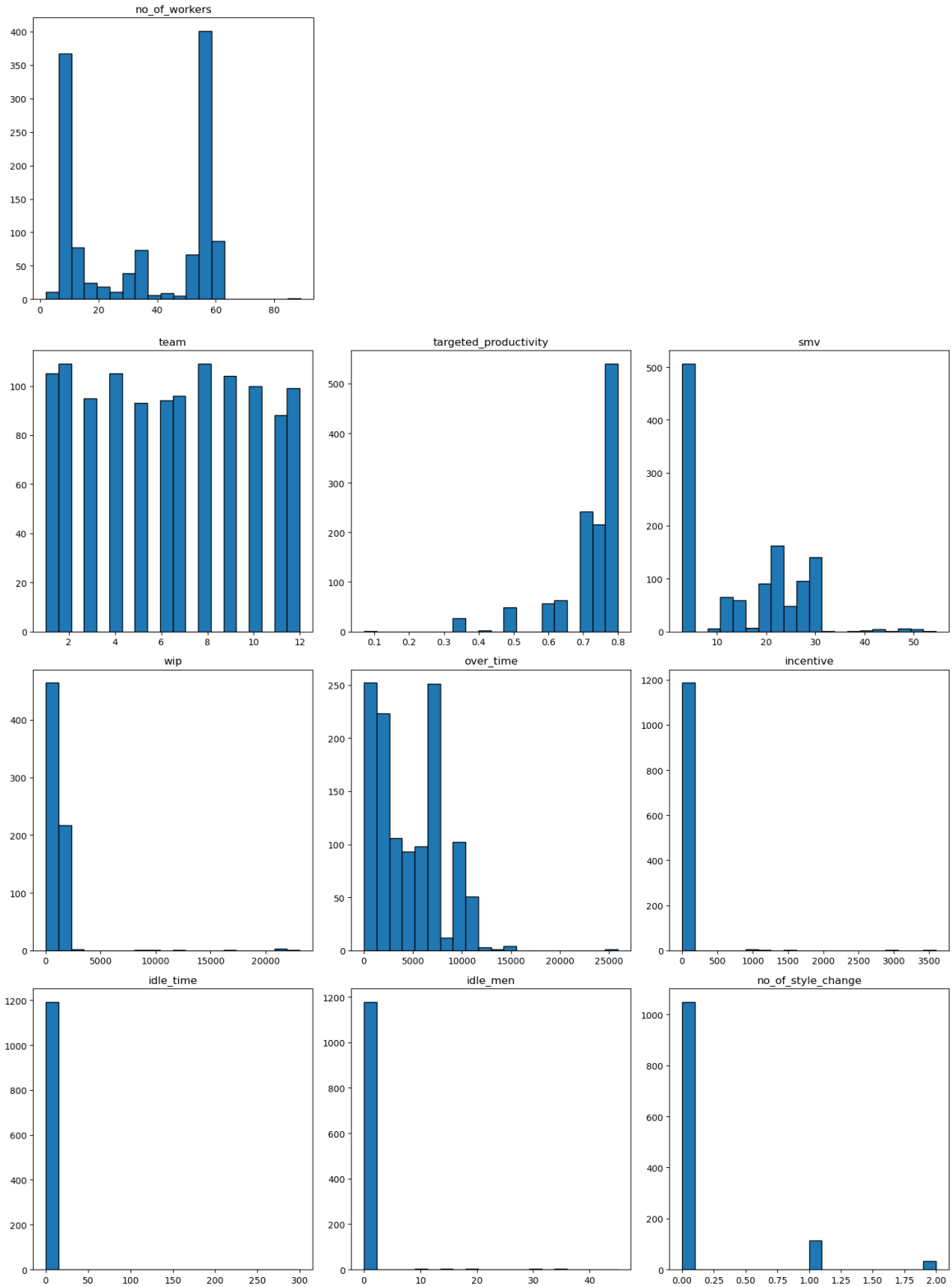


- Feature Relationships: A correlation heatmap showed that operational variables such as incentive and over_time had stronger associations with productivity compared to calendar-based variables.



The target variable was well-behaved, with most observations in the expected productivity range. However, skewness and outliers suggested that models must account for irregular cases. The large proportion of missing values in wip required careful imputation or exclusion.

Here are the other related plots.



Data Preprocessing

The preprocessing pipeline was designed to ensure numerical stability, preserve categorical structure, and capture non-linear interactions. The following steps were applied:

1. Datetime Transformation
    - The date column was converted to a proper datetime format.
    - Derived temporal features were added:
        - month (month of production).
        - weekday (day of the week).
2. Feature Engineering
    - Overtime per Worker: over_time divided by no_of_workers, adjusted with a small constant to avoid division by zero.
    - Idle Ratio: idle_men divided by no_of_workers, also adjusted for stability. These ratios normalized workload and idle metrics relative to workforce size.
3. Feature Groups
    - Categorical: quarter, department, day, team
    - Numerical (Linear): wip, over_time, incentive, idle_time, idle_men, no_of_style_change, month, weekday
    - Numerical (Polynomial Expansion): targeted_productivity, smv, no_of_workers, over_time_per_worker, idle_ratio
4. Transformations Applied
    - Linear Numerical Features:
        - Median imputation for missing values, with missing-indicator flags.
        - Standard scaling to zero mean and unit variance.
    - Polynomial Numerical Features:
        - Median imputation.
        - Standardization before and after polynomial expansion.
        - Pairwise interaction terms generated (degree = 2, interaction only, no bias term).
    - Categorical Features:
        - Most-frequent imputation for missing categories.
        - One-hot encoding with first-category drop to reduce multicollinearity.
5. Final Assembly
    - The transformations were combined via a ColumnTransformer.
    - Training and test sets were preprocessed separately to prevent data leakage:
        - Training set: X_tr_pre
        - Test set: X_te_pre

This preprocessing ensured that categorical data was fully usable in linear models, numerical variables were normalized, and key interaction effects were captured through polynomial expansion. The approach balanced model interpretability with flexibility,

allowing both linear and non-linear relationships with actual_productivity to be modeled effectively.

Model Development

Model 1: SGD Regression

Hyperparameter tuning was performed using 3-fold cross-validation to optimize the SGDRegressor model:

Parameter	Optimal Value	Parameter	Optimal Value
Loss	Huber	Learning Rate	Optimal
Penalty	L2	Eta_0	0.001
Alpha	0.001	Max Iterations	8000
L1 Ratio	0.2	Epsilon	0.03

The Huber loss function was selected for robustness to outliers, with L2 regularization to prevent overfitting.

Model 2: OLS Regression

Ordinary Least Squares regression was implemented using statsmodels library, fitting a linear model with all 46 predictors plus intercept term. Here is the top half of the results table:

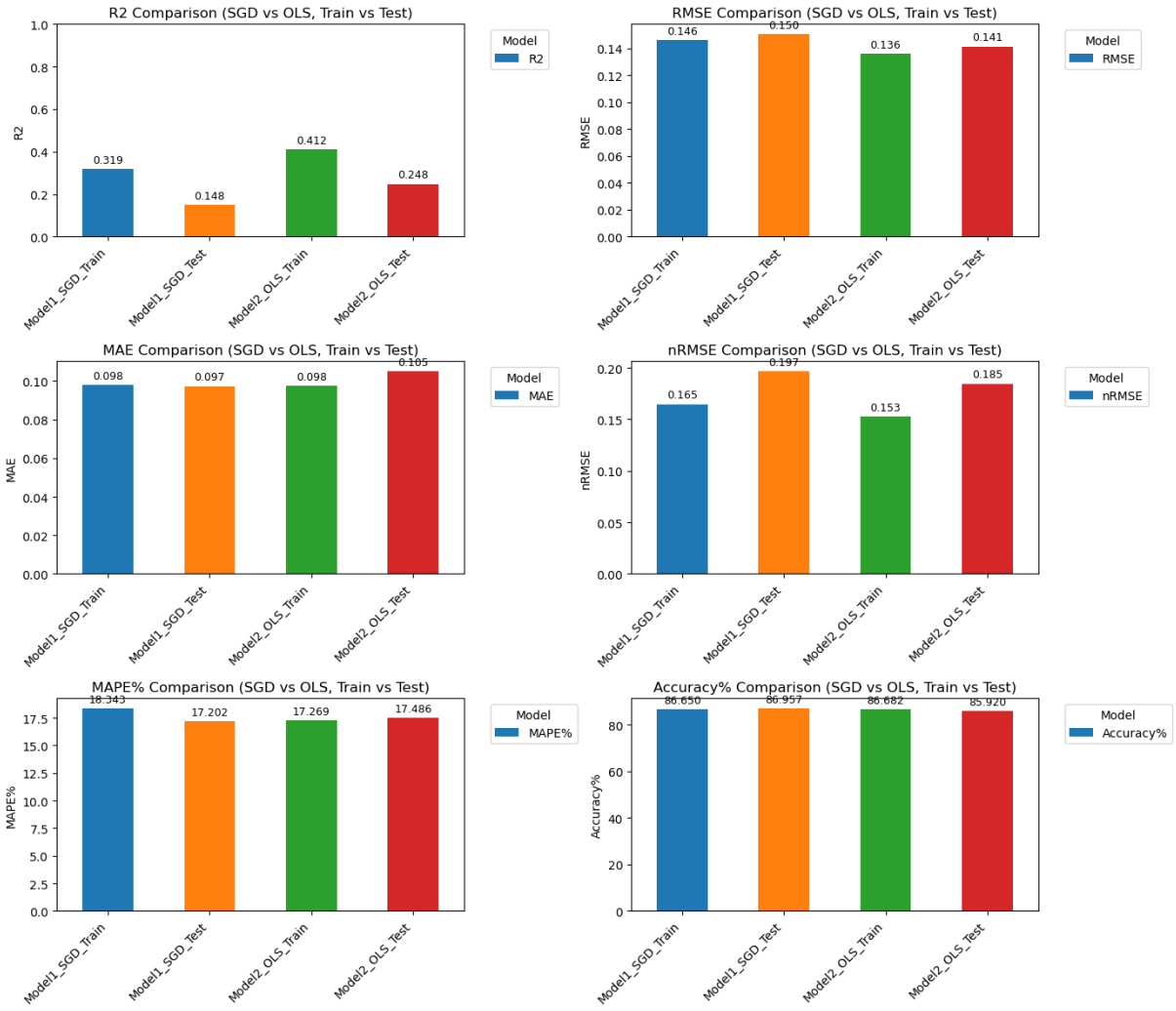
Metric	Value	Metric	Value
Dependent Variable	actual_productivity	Model	OLS
Method	Least Squares	Date	Sun, 21 Sep 2025
Time	20:18:12	No. Observations	957
Df Residuals	914	Df Model	42
R^2	0.412	Adj. R^2	0.385
F-statistic	15.28	Prob (F-statistic)	6.07e-79
Log-Likelihood	553.30	AIC	-1021.0
BIC	-811.5	Covariance Type	nonrobust

Results Analysis

Performance Comparison

Metric	SGD Train	SGD Test	OLS Train	OLS Test
R^2	0.319	*0.148*	0.412	*0.248*
RMSE	0.146	*0.150*	0.136	*0.141*
MAE	0.098	*0.097*	0.098	*0.105*

nRMSE	0.165	*0.197*	0.153	*0.185*
MAPE%	18.34	*17.20*	17.27	*17.49*
Accuracy%	86.65	*86.96*	86.68	*85.92*



Model Performance Analysis

- R^2 Analysis: The OLS model shows superior performance with test R^2 of 0.248 compared to SGD's 0.148, indicating OLS explains 67% more variance in the test data. Both models exhibit overfitting, with training R^2 substantially higher than test R^2 (OLS: 0.412 vs 0.248; SGD: 0.319 vs 0.148).
- Error Metrics: OLS demonstrates lower RMSE values (0.141 test) compared to SGD (0.150 test), suggesting better overall prediction accuracy. MAE values are similar between models (~0.10), indicating comparable absolute error magnitudes. The nRMSE values show OLS has better normalized error performance.
- Percentage Metrics: MAPE values are consistent across models (17-18%), while accuracy percentages remain high (85-87%) for both approaches, suggesting reasonable predictive capability despite moderate R^2 values.

Statistical Analysis (OLS Model)

Overall Model Significance:

- F-statistic: 15.28 with p-value = 6.07e-79 (highly significant)
- R^2: 0.412 (explains 41.2% of training variance)
- Adjusted R^2: 0.385 (accounts for number of predictors)

Model Comparison and Interpretation

- SGD vs OLS Performance: The OLS model consistently outperforms SGD across most metrics, particularly in R^2 values. This suggests that the linear relationships in the data are better captured through direct least squares estimation rather than gradient-based optimization with the chosen hyperparameters.
- Overfitting Assessment: Both models show evidence of overfitting, with substantial drops from training to test performance. The OLS model maintains better generalization despite having more parameters.

Conclusions

The comparative analysis demonstrates that OLS regression outperforms SGD regression for this dataset, achieving superior predictive accuracy and lower error rates. However, both models face limitations including overfitting tendencies and moderate overall explanatory power (best test $R^2 = 0.248$). This is primarily low Standard Deviation, making the model estimates close to the overall mean. Due to this, both models are very limited by the variability of the data. Both models achieve reasonable practical accuracy (85-87%) despite the unexplained variability in worker productivity that may require additional features or non-linear modeling approaches.