INDENG 142A Final Report

Introduction:

The semiconductor industry is extremely sensitive to fluctuations in manufacturing capacity, and utilization rates above 90% result in significantly longer lead times and major financial losses for downstream manufacturers. The recent chip shortages in 2021 and 2022 highlighted the severe consequences of such stresses in capacity planning, as automakers were forced to halt production lines and market variability surged. The motivation of this project is to forecast semiconductor fab capacity utilization 1 to 3 months in advance, enabling companies to make proactive procurement and production planning decisions. By leveraging historical production, capacity, and automotive demand data, our machine learning models aim to anticipate periods of high utilization before they create bottlenecks for companies. Ultimately, accurate predictions from our models will help mitigate supply chain disruptions, reduce financial losses, and support data-driven decision-making for one of the most innovative industries in the world.

Data Collection:

The data from this project is collected from FRED (Federal Reserve Economic Data), a significant repository of U.S. economic and industrial statistics for different industries. The target variable is *semiconductor capacity utilization*, and is measured monthly from January 1972 to the present, providing over 636 observations. This series reflects the percentage of manufacturing capacity currently used at semiconductor fabs, making it a crucial indicator of semiconductor supply chain stress.

The predictor variables below are also collected from FRED:

- *Semiconductor industrial production index (IPG):* Captures total semiconductor output and serves as a supply-side indicator.
- *U.S. domestic auto production (DAUP):* Acts as a demand-side indicator, reflecting downstream pressure from the automotive industry, which has lead times of 3–6 months.

Then, to better capture trends and relationships, we derive several features based on the data collected above:

1. *CAPUTL_lag1 (Capacity utilization 1 month ago)*: Exploits the high autocorrelation in capacity utilization, helping predict near-future utilization based on the previous month.
2. *CAPUTL_lag3 (Capacity utilization 3 months ago)*: Captures medium-term trends, distinguishing temporary spikes from sustained high utilization periods.
3. *IPG_lag1 (Production index 1 month ago)*: Reflects the direct effect of recent production on current capacity usage.
4. *IPG_lag3 (Production index 3 months ago)*: Captures quarter-level trends in production that affect future capacity stress.
5. *DAUP_lag3 (Auto production 3 months ago)*: Accounts for automotive supply chain lead times, linking past auto demand to current semiconductor utilization.

6. *CAPUTL_ma3 (3-month rolling average of capacity utilization)*: Smooths short-term volatility to highlight trends more clearly.
7. *IPG_ma3 (3-month rolling average of production index)*: Reduces noise from seasonal fluctuations, providing a clearer signal of underlying production trends.
8. *Month (1–12)*: Provides an insight regarding the seasonality of semiconductor demand
9. *Quarter (1–4)*: Captures the business cycle patterns and measurements in semiconductor manufacturing and BU planning
10. *IPG_x_DAUP (Production index × Auto production for current month)*: Models combined supply-demand stress, identifying periods when high production coincides with strong downstream demand.

Our engineered features ensure that the dataset captures both short-term and long-term trends, seasonal patterns, and interactions between supply and demand, ultimately resulting in a robust foundation for analytically forecasting semiconductor capacity utilization.

Model 1: Multiple Linear Regression

Using the features above, we created a Multiple Linear Regression model to predict the semiconductor capacity utilization from the other features. We split the dataset into training and testing sets, such that the training set included all data before January 2020, and the testing set included all data from January 2020 and onwards. Then, we calculated the Variance Inflation Factors (VIF) for each of the features, and iteratively removed all of the variables with high multicollinearity in order to derive our final features for the model. These features are given below:

final_features = ["CAPUTL_lag1", "CAPUTL_lag3", "IPG_lag3", "DAUP_lag3", "Month", "IPG_x_DAUP"]

Then, from this, we formulated an ordinary least squares (OLS) regression model through the use of the statsmodels library, using the following formula for the model:

$$CAPUTL_t = 3.8956 + 1.2993 \cdot CAPUTL_{t-1} - 0.3527 \cdot CAPUTL_{t-3} + 0.0134 \cdot IPG_{t-3} + 0.0014 \cdot DAUP_{t-3} - 0.0260 \cdot Month_t - 0.00005257 \cdot (IPG \cdot DAUP)_t$$

From this, the R^2 score of the model comes out to be 0.978, which indicates that the model is a strong fit for predicting the capacity utilization based on the given features. And the MAE and RMSE for the model on the test set come out to be 1.333 and 1.583 respectively.

Model 2: Random Forest Regressor

We also incorporated a Random Forest model to predict semiconductor capacity utilization from the other features. Similar to the previous model, we split the dataset to incorporate all data before 2020 as the training set, and all data from 2020 and onwards as the testing set. Then, we set different parameter values for our model as follows: n_estimators: (100, 250, 500), max_depth: (None, 5, 10, 20), min_samples_split: (2, 5, 10), and min_samples_leaf: (1, 2, 4).

Then, we also use cross-validation in this model using a TimeSeriesSplit with 5 splits, which ensures that no future data values leak into our training dataset, and GridSearchCV tests every combination of our hyperparameters across these splits, and scores with a negative MAE, meaning that a lower MAE value means that the parameter is better. From this, GridSearchCV gives us the optimal values of our hyperparameters as max_depth=10, min_samples_leaf=4, min_samples_split=5, and n_estimators=250. From this, we then use these hyperparameter values to train the Random Forest regression model, and through these parameter values, we enable the model to capture non-linear patterns, lag effects, seasonality, and interactions. From this, the code outputs the MAE and RMSE of the model when performed on the test set, which come out to be 0.922 and 1.138 respectively.

Model 3: XGBoost Regressor

Similar to the previous model, this model splits the dataset and incorporates all data before 2020 in the training set, and all data from 2020 and onwards into the testing set. Then, we also remove any missing values in the dataset. This model uses all 10 features for predicting the semiconductor capacity utilization variable. This model allows us to handle non-linear relationships between the variables, and is robust to collinearity. The parameters selected for this model are as follows: learning_rate: [0.05, 0.1, 0.2], n_estimators: [50, 100], and max_depth: [3, 5]. These parameters allow for a robust model with slower and more stable learning patterns. Furthermore, the model also uses TimeSeriesSplit for cross-validation, with three splits. This prevents leakage of future values into the model's predictions, preserves temporal order, and also allows the model to make more realistic predictions based on the data. After this, we calculate the MAE and RMSE for this model, which come out to be 0.9961 and 1.2377 respectively. Finally, we also calculate the importance of each feature for the model's predictions. Upon doing so, CAPUTL_lag1 and CAPUTL_ma3 are the two features which have the most importance for this model's predictions, and by a large margin compared to the other features.

Model Comparisons:

In the table below, we output the MAE and RMSE values for each of the four models, and use the table to determine which model has the best performance on the testing dataset:

Model	MAE Value	RMSE Value
Multiple Linear Regression	1.333	1.583
Random Forest Regressor	0.922	1.138
XGBoost Regressor	0.9961	1.2377

From this, since both the MAE and RMSE values for the Random Forest regressor model are smaller than for any of the other two models, we can conclude that the **Random Forest Regressor model has the strongest performance on the testing set out of all the three models.**

Conclusion:

Using data from the FRED database, which included lagged utilization values, production indicators, and seasonality features, this project evaluated three different models to predict semiconductor capacity utilization. Based on the performance results of the three models, Random Forest has the strongest performance, achieving a test MAE of 0.92 and a RMSE value of 1.14. While the other models performed moderately well, the Random Forest model exhibited more stability and accuracy overall. With such an exceptional performance, the results of the Random Forest model indicate the strong ability to identify when capacity utilization is likely to exceed the 90% threshold. Despite Random Forest performing exceptionally well, as with the other models' performances, there exist limitations. Random Forest and the other models evaluated, used U.S. aggregate data, which could potentially mask or overlook firm-level or regional differences within the semiconductor supply chain. Moreover, the predictive power of the models is driven by recent capacity utilization trends. That being said, none of the models can serve as the sole predictor of future capacity utilization, as they cannot predict unprecedented disruptions, similar to pandemic-related shutdowns or other events. For future work, it would be wise to consider incorporating firm-level production or different regional data to capture variations across the industry.

However, thinking practically, the Random Forest model and its performance have value. The accuracy it demonstrated could provide great aid and support for CFO and supply chain managers. It is of immense importance, though, to recognize that overreliance on the model could lead to overinvestment and, consequently, exacerbate shortages and market instability.

Comprehensively, the impact of this project is to provide actionable, data-driven forecasts of semiconductor capacity utilization, enabling companies to anticipate periods of high demand and mitigate supply chain disruptions. By evaluating multiple models, the project demonstrated that the Random Forest Regressor model achieves the strongest predictive accuracy, and is robustly able to capture trends, seasonality, and interactions between supply and demand. These results offer practical support for operational and financial decision-making in the semiconductor industry, helping reduce losses and improve planning efficiency.

Link To Data Sources:

Semiconductor capacity utilization, monthly (Jan 1972 - Present, 636+ observations)
https://fred.stlouisfed.org/series/CAPUTLG3344S
Semiconductor industrial production index (supply-side indicator)
https://fred.stlouisfed.org/series/IPG3344S\
U.S. domestic auto production, seasonally adjusted (demand-side indicator)
https://fred.stlouisfed.org/series/DAUPSA

Link To Code And Output Results:
https://github.com/KevinZWong/INDENG142a_Project

Additional Figures:
https://drive.google.com/drive/folders/1c03Al1kW6NPSoP8vyImkh-56XS8TeJci?usp=sharing