Some information on specific columns:

CLNDR_DT = revenue transaction date

CLNDR_DT_MONTHSEQ = sequential month number that corresponds with CLNDR_DT

Auth_Date = The date from which a project is considered new. Because we do 12 month forecast, we look at this date +12 months as the date frame from where a project is new. For example if we are doing a forecast for January 2023- December 2023, if a project has an auth date of January 2023, we would not have historical data from <December 2022 so any revenue generated by this project would be considered 'new' for 2023.

Auth_Date_Monthseq = sequential month number that corresponds with Auth_Date

Question:

How would you approach predicting what NEW PROJECT revenue will be for the next 12 months at the CATEGORY level. As noted above there will be no history for a new project. However, there are histories of projects that were considered 'new' from a particular point in time. The attached dataset was raw export with the data not filtered or cleaned up in anyway. Please review the data to be able to provide some specific details in answering the questions below.

- How would you think creating a 'new project' model? What type of approaches would you try?
-

I utilized time series models based on past data to create average expense estimates by category (a total of 7 categories). Each new project is assigned a category, and using historical project data, we provide an average estimate of where their values might fall. Due to time constraints, we have focused only on Area1, but a thorough approach for estimating new projects by their area would involve this analysis for all categories. The table below will be used as input for forecasting both the year 2023 (filtering out the existing data up to May 2024) and for predictions from June 2024 to July 2025. We used machine learning models, specifically RandomForest, to carry out this analysis.

- How would you transform this dataset so that it could be used for the approach you chose above? In this way

	index	CLNDR...	Revenue	CLNDR...	Auth_D...	Auth_D...	Category	project	Month	Year	TS
0	0	2023-06-...	20270.25	174	2023-06-...	174	Area3	Project118...	6	2023	2023-06
1	1	2023-07-...	4552.99	175	2023-06-...	174	Area3	Project118...	7	2023	2023-07
2	2	2023-08-...	0	176	2023-06-...	174	Area3	Project118...	8	2023	2023-08
3	3	2023-09-...	0	177	2023-06-...	174	Area3	Project118...	9	2023	2023-09
4	4	2023-10-3...	0	178	2023-06-...	174	Area3	Project118...	10	2023	2023-10
5	5	2023-11-3...	1579.5264...	179	2023-06-...	174	Area3	Project118...	11	2023	2023-11
6	6	2023-12-3...	8786.683...	180	2023-06-...	174	Area3	Project118...	12	2023	2023-12
7	7	2024-01-3...	459.19	181	2023-06-...	174	Area3	Project118...	1	2024	2024-01
8	8	2024-02-...	199.03115...	182	2023-06-...	174	Area3	Project118...	2	2024	2024-02
9	9	2024-03-...	0	183	2023-06-...	174	Area3	Project118...	3	2024	2024-03
10	10	2023-02-...	0	170	2023-02-...	170	Area1	Project194	2	2023	2023-02
11	11	2023-03-...	0	171	2023-02-...	170	Area1	Project194	3	2023	2023-03
12	12	2023-04-...	378.11552...	172	2023-02-...	170	Area1	Project194	4	2023	2023-04
13	13	2023-05-...	-73.00305...	173	2023-02-...	170	Area1	Project194	5	2023	2023-05
14	14	2023-06-...	-0.006208...	174	2023-02-...	170	Area1	Project194	6	2023	2023-06
15	15	2023-07-...	1830.5849...	175	2023-02-...	170	Area1	Project194	7	2023	2023-07
16	16	2023-08-...	910.24066...	176	2023-02-...	170	Area1	Project194	8	2023	2023-08
17	17	2023-09-...	-235.84	177	2023-02-...	170	Area1	Project194	9	2023	2023-09
18	18	2023-10-3...	0.01	178	2023-02-...	170	Area1	Project194	10	2023	2023-10
19	19	2023-11-3...	0	179	2023-02-...	170	Area1	Project194	11	2023	2023-11
20	20	2023-09-...	0	177	2023-09-...	177	Area4	Project124...	9	2023	2023-09
21	21	2023-10-3...	0	178	2023-09-...	177	Area4	Project124...	10	2023	2023-10
22	22	2023-11-3...	0	179	2023-09-...	177	Area4	Project124...	11	2023	2023-11
23	23	2023-12-3...	0	180	2023-09-...	177	Area4	Project124...	12	2023	2023-12
24	24	2024-01-3...	115207.49	181	2023-09-...	177	Area4	Project124...	1	2024	2024-01
25	25	2024-02-...	342104.64	182	2023-09-...	177	Area4	Project124...	2	2024	2024-02

- How would you filter data, would you create any new variables, etc? What would your test and training datasets look like? -→ **Please see the Jupyter Notebook**

I took the following steps:

1. **Data Filtering**:
   - We filtered the dataset to include data from January 2018 to June 2023 for training, and from July 2023 to May 2024 for validation. This ensures the model is trained on sufficient historical data and validated before making future predictions.
2. **New Variables**:
   - We created new time-based features such as `Year` and `Month` to capture temporal patterns in the data.
   - Additionally, a combined feature `TS` representing the year and month in the format 'YYYY-MM' was created for easier time series analysis.
3. **Training and Test Datasets**:
   - The training dataset included data from January 2018 to June 2023.
   - The validation dataset included data from July 2023 to May 2024.
   - For future predictions, we extended the dataset to include the next 14 months (June 2024 to July 2025).
4. **Machine Learning Models**:
   - We used RandomForest, a machine learning model, to perform the forecasting. The model was trained using the training dataset and validated using the validation dataset. Predictions were then made for the future period from June 2024 to July 2025.

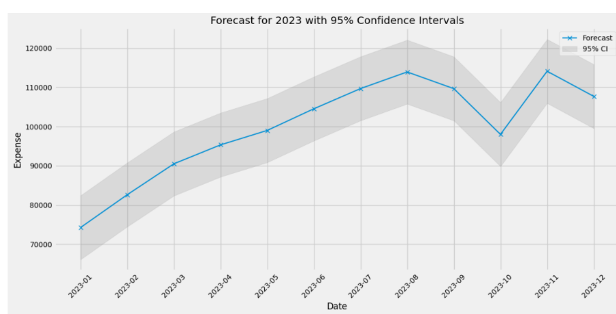
- Please provide examples of what these inputs into the model would be for (variables: dependent, independent, static, dynamic):

Variables:

1. **Dependent Variable**:
    - **`av_expense_cate1`**: This is the target variable we aim to forecast, representing the average expenses for Area1.
2. **Independent Variables**:
    - **`Year`**: Captures the yearly trend in the data.
    - **`Month`**: Captures the monthly seasonality in the data.
    - **`TS`**: A combined feature representing the year and month in the format 'YYYY-MM', useful for time series analysis.
3. **Static Variables**:
    - **`Category`**: The area or category to which the project belongs (in this case, `Area1`). This variable does not change over time and helps in grouping projects by their category.
4. **Dynamic Variables**:
    - **`Year`**: Changes over time and captures the annual trend.
    - **`Month`**: Changes monthly and captures seasonality effects.

- Predicting from the beginning of 2023 (forecast for all new project revenue for full year (months 1-12).
-

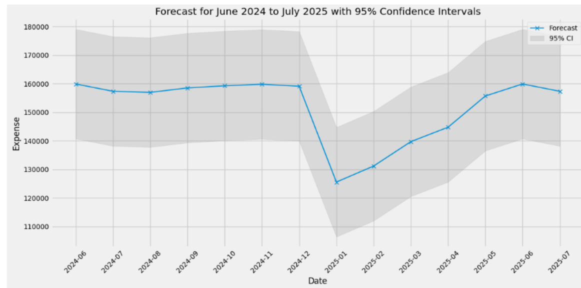
	TS	Forecast	Lower_CI	Upper_CI
0	2023-01	74265.585500	66113.816017	82417.354983
1	2023-02	82621.712250	74469.942767	90773.481733
2	2023-03	90502.789325	82351.019842	98654.558808
3	2023-04	95325.520575	87173.751092	103477.290058
4	2023-05	99001.184300	90849.414817	107152.953783
5	2023-06	104523.592000	96371.822517	112675.361483
6	2023-07	109664.484575	101512.715092	117816.254058
7	2023-08	113908.045692	105756.276208	122059.815175
8	2023-09	109651.797292	101500.027808	117803.566775
9	2023-10	97994.557092	89842.787608	106146.326575
10	2023-11	114090.678933	105938.909450	122242.448417
11	2023-12	107682.627833	99530.858350	115834.397317



- Would your datasets look any different for predicting from July 2024 to July 2025?

  YES, revenues values changed over the time

```
Best parameters for RandomForest: {'max_depth': 10, 'min_samples_split':
         TS       Forecast       Lower_CI       Upper_CI
0   2024-06  159918.830946  140768.987288  179068.674604
1   2024-07  157356.579223  138206.735565  176506.422881
2   2024-08  156974.946083  137825.102425  176124.789741
3   2024-09  158533.638058  139383.794400  177683.481716
4   2024-10  159279.430467  140129.586809  178429.274125
5   2024-11  159813.855456  140664.011798  178963.699114
6   2024-12  159133.189282  139983.345624  178283.032940
7   2025-01  125597.361806  106447.518148  144747.205464
8   2025-02  131203.603325  112053.759667  150353.446983
9   2025-03  139738.932216  120589.088558  158888.775874
10  2025-04  144788.206340  125638.362682  163938.049998
11  2025-05  155696.860707  136547.017049  174846.704365
12  2025-06  159918.830946  140768.987288  179068.674604
13  2025-07  157356.579223  138206.735565  176506.422881
```



Conclusion

In this analysis, I aimed to forecast expenses for new projects based on historical data. The steps we followed included:

1. **Data Preprocessing**: We filtered the dataset to include data from January 2018 to June 2023 for training, and from July 2023 to May 2024 for validation.
2. **Feature Engineering**: We created new time-based features such as `Year`, `Month`, and a combined feature `TS` representing the year and month in the format 'YYYY-MM'.
3. **Model Selection and Training**: We selected the RandomForest model for its robustness and applied hyperparameter tuning using GridSearchCV. The model was trained on the training dataset and validated using the validation dataset.
4. **Forecasting**: We generated forecasts for the period from June 2024 to July 2025, including confidence intervals to capture the uncertainty of predictions.
5. **Evaluation**: We evaluated the model's performance using metrics like MAD, MSE, AIC, BIC, and R^2, and validated it through a series of tests.

Limitations and Future Work

Due to time constraints, we limited our model selection to a few classic time series and machine learning models. In a comprehensive analysis, additional models such as ARIMA, ETS, Gradient Boosting Machines, and others should be considered to ensure robust forecasting. Deep learning models were not employed due to the limited number of data points (less than 100 periods), which is typically insufficient for effective deep learning model training.

Next Steps

1. **Deployment**: After finalizing the model selection, the next step would be to deploy the model in a production environment. This involves setting up a pipeline for continuous data integration and model retraining as new data becomes available.
2. **Monitoring and Validation**: Once deployed, it's crucial to monitor the model's performance to ensure it continues to provide accurate forecasts. Techniques such as A/B testing can be used to compare the performance of different models or scenarios.
3. **Performance Testing**: Simulating various scenarios and stress-testing the model can help identify potential weaknesses and areas for improvement. This involves testing the model under different conditions to see how well it generalizes to unseen data.

By following these steps, we can ensure that the forecasting model remains accurate and reliable, providing valuable insights for planning and decision-making in new projects.