Data Warehouse Assignment

Case Study

April 2024

Project Description

A company called ForestAI is mainly engaged in applying Data Science to perform predictions of forest inventory. Its AI platform utilizes climate, geo, and customer process data to provide efficient identification of potential purchasing and harvestable areas in forests with less field visits, helping make more informed and timely pricing decisions.

Performing quality validation of its AI software is a critical concern of ForestAI. The core of any model validation is performance testing by conducting some benchmark, such as providing comparisons of the new model's outputs to those of the version of the model it is replacing, comparing the model being validated to some other model or metrics, or comparing the new model's outputs to those of an external "challenger" model which works towards the same objective or utilizes the same data. Choosing what kind of benchmark to use within a model validation can sometimes be a difficult task and depends on the type of model being tested.

For each client, the company conducts benchmarks to validate models, compare models outputs and the ground truth's outputs, and choose one model to integrate it into the product. ForestAI also needs on occasion to show these comparisons to the clients in order to build its clients' trust in the product. However, the pipeline of validating, choosing, and presenting results to clients depends on each client. It is non-standard and non reusable. Consequently, members of the data science team produce different pipelines, use different coding tools, and present results in multiple platforms. The pipelines need to be re-exectued every time there is a need to conduct a benchmark. Moreover, model validation is incomplete without analyzing these benchmarks, which is an integral part of the process. The quality assurance team and managers analyze different types of benchmarks, but struggle to find the results within the platform. These results are also presented in tables, images. and excel sheets, which is not ideal for supporting decision-making. The platform thus suffers from poor user experience, limited visualization, and a non standardization of data pipelines, all of which significantly slows down the model quality validation process.

You were hired to help overcome the limitation of the current platform by improving the benchmarking process to fit the company needs. This should be done by creating one pipeline that takes data from all disparate sources (ground truth, company's predictions, clients models, competitors models..) into one common destination, enabling quick data analysis for technical and business insights, ensuring consistent data quality, which is absolutely crucial for reliable business insights, and adding intermediate storage to improve performance.

The aim is thereby to redesign and standardize the data pipeline by identifying KPIs and implementing a data warehouse. And secondly, by creating interactive dashboards and storing them in one place to improve clients' user experience.

Project Specifications

ForestAI's AI solution uses weather information from many years back, soil classifications, optical and other images taken from satellites, LIDAR information, etc. to identify and model the parameters that influence growth of trees, distribution of species, and other characteristics of forest inventory. The output is a prediction of wood mass by species and other parameters critical to forest inventories. Here is an example of predictions targets for certain clients from the AI application:

- **Total volumes** by species (pine, birch, spruces, deciduous.) in an area.
- Trees **height** and **diameter**.
- **Basal area**: an index that tells forest managers how dense or crowded a forest is. Forest managers need to know how much competitions and crowdedness there is among trees.
- **Woods volumes** to help buyers know what they can expect from some types of species, for example, to define if the wood will be used as household firewood, or to produce papers and boards...

After building any model for a client, predictions targets need to be checked and validated by comparing different versions of model's outputs to the **ground truth data** and **client's estimations**. This helps prove that ForestAI's models are performing well.

Customer estimations come from forest inventory. It is a systematic collection of data and forest information for assessment or analysis. Forest managers generalizes these estimations for bigger areas based on their experiences, which may be wrong, that's why some forest managers want to check if these estimations can be used for managing forests. The goal of ForestAI is to build models with results better than foresters', therefore, ForestAI compare its predictions to these data to convince clients to buy its product.

Ground truth come from real observations using approaches such as HPR (Harvested Production Report), LIDAR, and Sample plots. Clients provide these data to ForestAI since they are forest managers and investors. Therefore, the type of the ground truth differs among clients.

Testing and validating model quality refers to validating the quality of the model created with a machine learning approach from some training data. Model quality is important when one needs an initial assessment of a model before going to production or if one wants to observe relative improvements from learning efforts or compare two models.

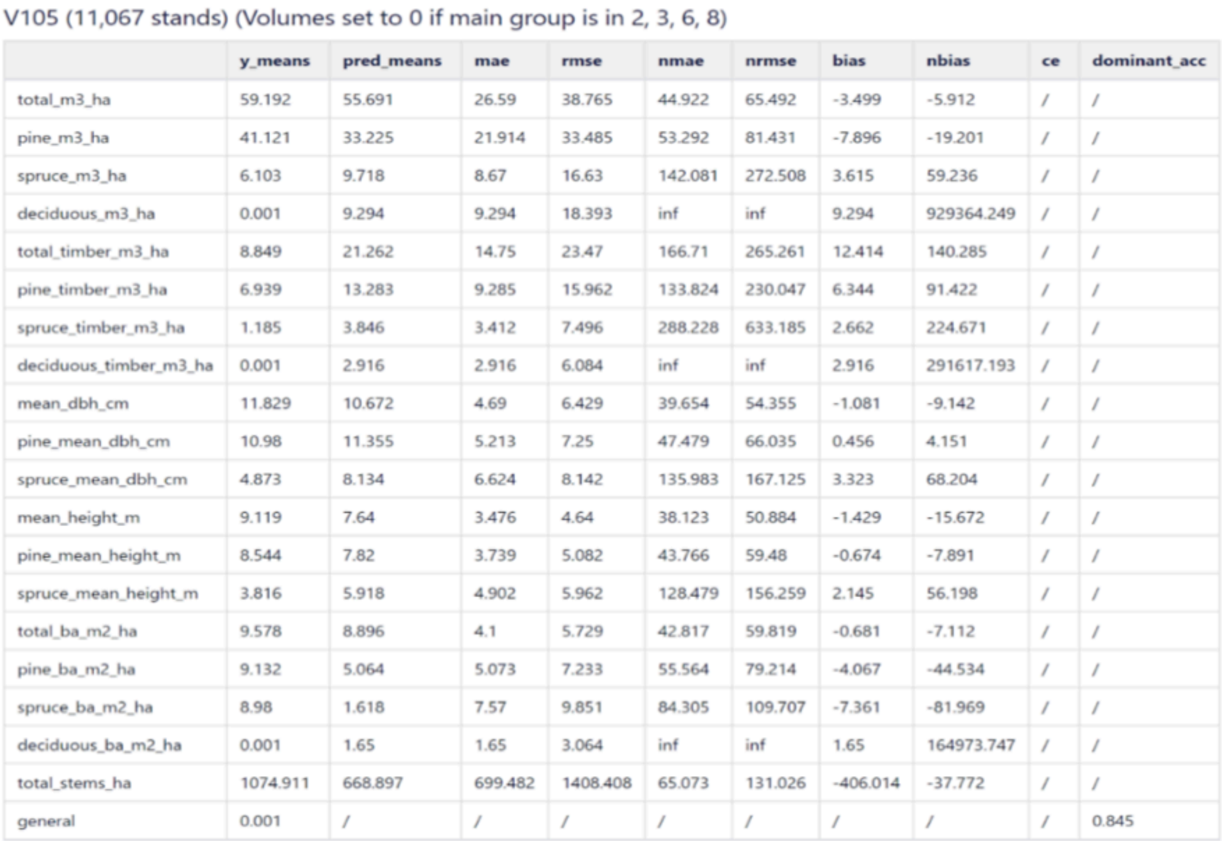
The main core of any model validation is performance testing. Benchmarking is one of the main performance testing components. It is a process where the validator is providing a comparison of the model at hand to some other model or metric. Benchmarking takes many forms and sometimes entails comparing the model's outputs to:

- The model's previous version
- An externally produced model
- A model built by the validator
- Other models and methodologies considered by the model developers, but not chosen
- Industry best practice

Additionally, statistical metrics are needed to calculate errors.

Existing Reports

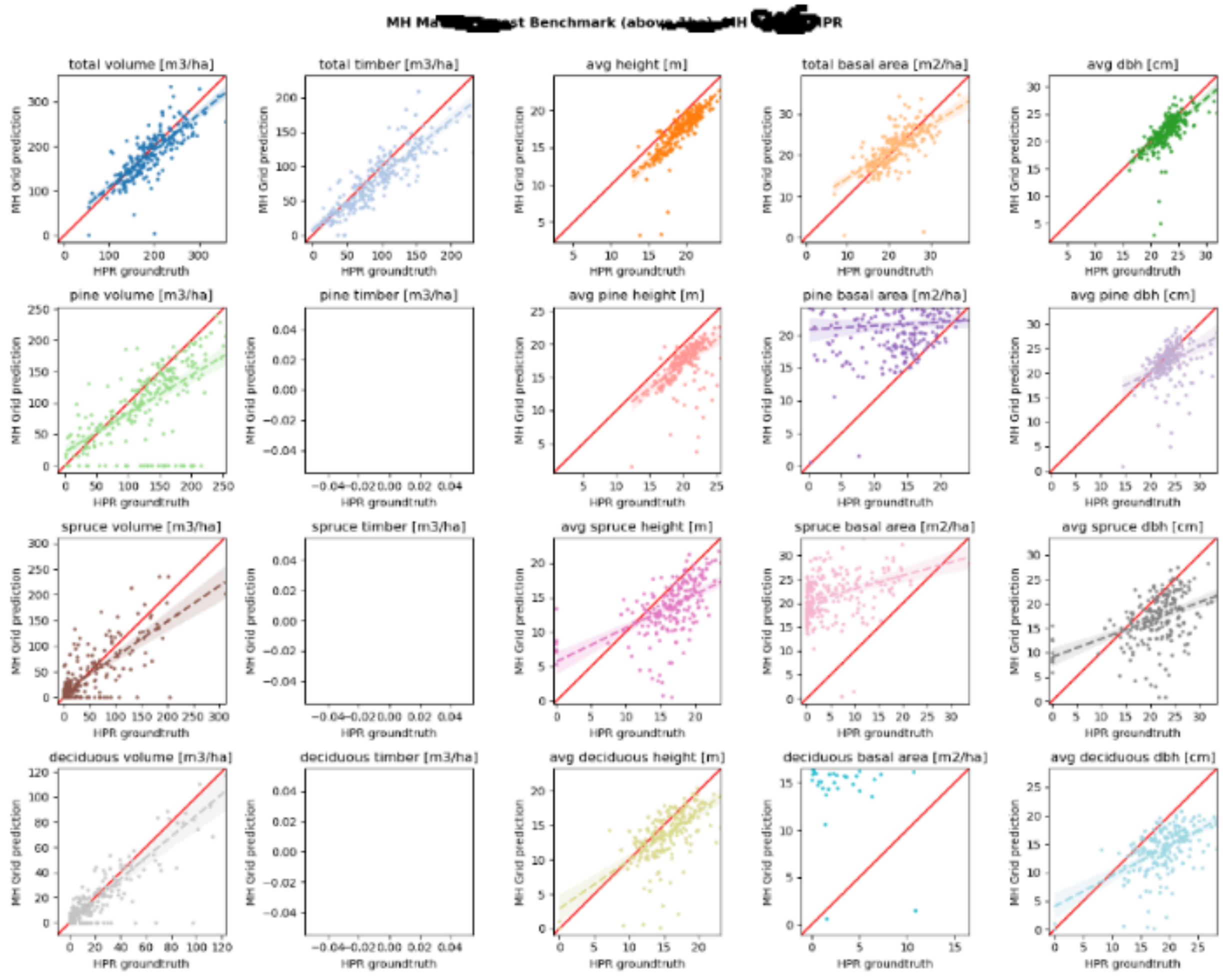
Model Performance Reports present characteristics and error metrics for the version of a model. The following example shows the performance metrics of version 105 of a prediction model:



V105 (11,067 stands) (Volumes set to 0 if main group is in 2, 3, 6, 8)

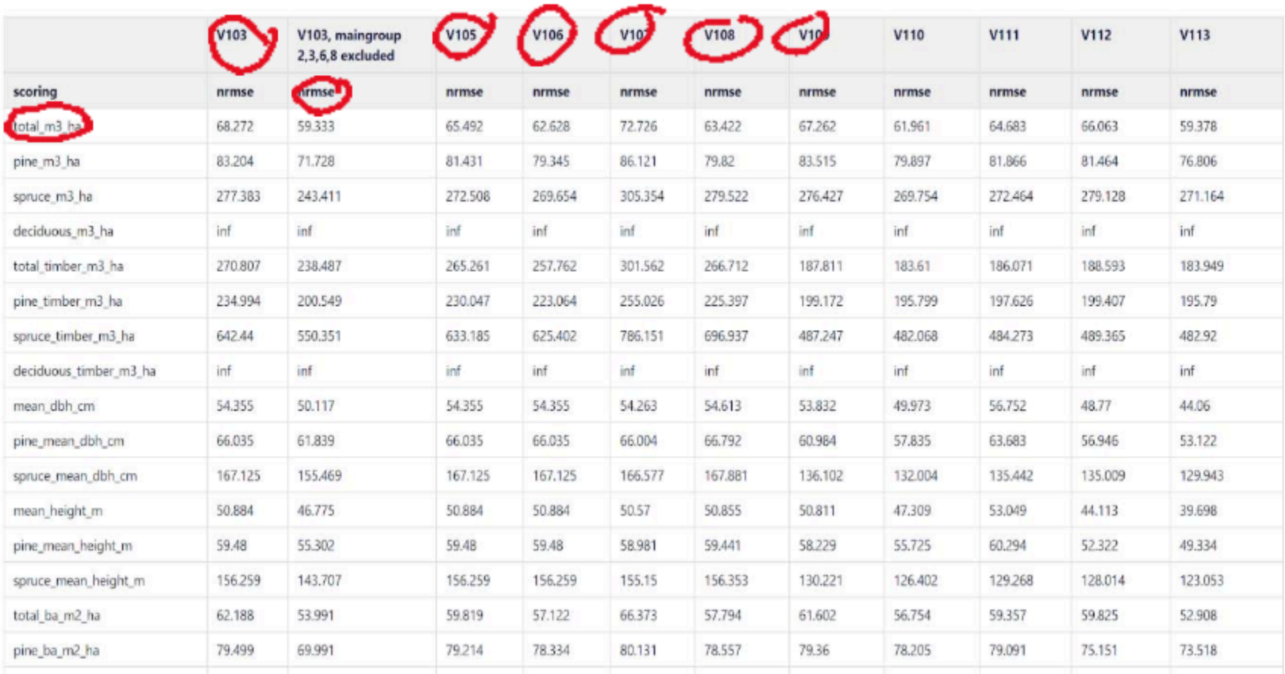
	y_means	pred_means	mae	rmse	nmae	nrmse	bias	nbias	ce	dominant_acc
total_m3_ha	59.192	55.691	26.59	38.765	44.922	65.492	-3.499	-5.912	/	/
pine_m3_ha	41.121	33.225	21.914	33.485	53.292	81.431	-7.896	-19.201	/	/
spruce_m3_ha	6.103	9.718	8.67	16.63	142.081	272.508	3.615	59.236	/	/
deciduous_m3_ha	0.001	9.294	9.294	18.393	inf	inf	9.294	929364.249	/	/
total_timber_m3_ha	8.849	21.262	14.75	23.47	166.71	265.261	12.414	140.285	/	/
pine_timber_m3_ha	6.939	13.283	9.285	15.962	133.824	230.047	6.344	91.422	/	/
spruce_timber_m3_ha	1.185	3.846	3.412	7.496	288.228	633.185	2.662	224.671	/	/
deciduous_timber_m3_ha	0.001	2.916	2.916	6.084	inf	inf	2.916	291617.193	/	/
mean_dbh_cm	11.829	10.672	4.69	6.429	39.654	54.355	-1.081	-9.142	/	/
pine_mean_dbh_cm	10.98	11.355	5.213	7.25	47.479	66.035	0.456	4.151	/	/
spruce_mean_dbh_cm	4.873	8.134	6.624	8.142	135.983	167.125	3.323	68.204	/	/
mean_height_m	9.119	7.64	3.476	4.64	38.123	50.884	-1.429	-15.672	/	/
pine_mean_height_m	8.544	7.82	3.739	5.082	43.766	59.48	-0.674	-7.891	/	/
spruce_mean_height_m	3.816	5.918	4.902	5.962	128.479	156.259	2.145	56.198	/	/
total_ba_m2_ha	9.578	8.896	4.1	5.729	42.817	59.819	-0.681	-7.112	/	/
pine_ba_m2_ha	9.132	5.064	5.073	7.233	55.564	79.214	-4.067	-44.534	/	/
spruce_ba_m2_ha	8.98	1.618	7.57	9.851	84.305	109.707	-7.361	-81.969	/	/
deciduous_ba_m2_ha	0.001	1.65	1.65	3.064	inf	inf	1.65	164973.747	/	/
total_stems_ha	1074.911	668.897	699.482	1408.408	65.073	131.026	-406.014	-37.772	/	/
general	0.001	/	/	/	/	/	/	/	/	0.845

Scatterplots are also used to compare the model's performance to ground truth data as shown in the example below:



Model Validation Reports compare a new model's outputs to those of the version of the model it is replacing. They are used to prove that the new model performs better by comparing all models to ground truth, to see whether the new model is both working as intended and accurate. The validator compares all models to the ground truth by calculating error metrics. Then chooses the model with the best error metrics.

The following figure shows a comparison between different versions of ML models (V103, V105, V106, V107.) and the HPR data as a ground truth by calculating first the NRMSE.

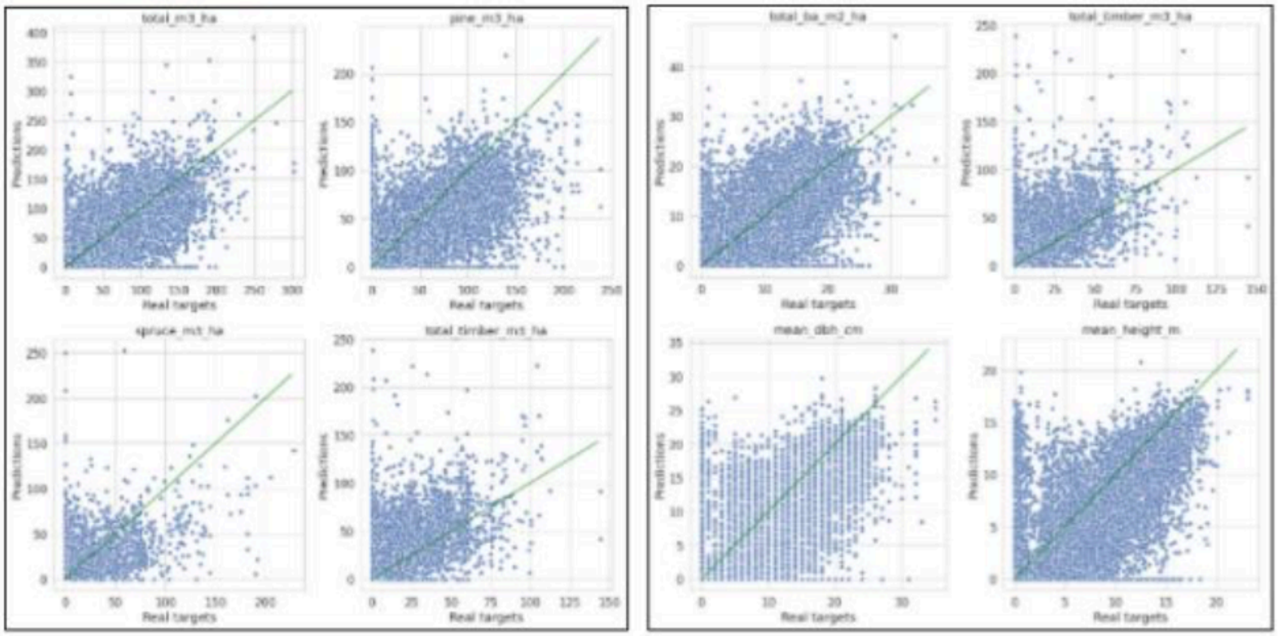


	V103	V103, maingroup 2,3,6,8 excluded	V105	V106	V107	V108	V109	V110	V111	V112	V113
scoring	nrmse	nrmse	nrmse	nrmse	nrmse	nrmse	nrmse	nrmse	nrmse	nrmse	nrmse
total_m3_ha	68.272	59.333	65.492	62.628	72.726	63.422	67.262	61.961	64.683	66.063	59.378
pine_m3_ha	83.204	71.728	81.431	79.345	86.121	79.82	83.515	79.897	81.866	81.464	76.806
spruce_m3_ha	277.383	243.411	272.508	269.654	305.354	279.522	276.427	269.754	272.464	279.128	271.164
deciduous_m3_ha	inf	inf	inf	inf	inf	inf	inf	inf	inf	inf	inf
total_timber_m3_ha	270.807	238.487	265.261	257.762	301.562	266.712	187.811	183.61	186.071	188.593	183.949
pine_timber_m3_ha	234.994	200.549	230.047	223.064	255.026	225.397	199.172	195.799	197.626	199.407	195.79
spruce_timber_m3_ha	642.44	550.351	633.185	625.402	786.151	696.937	487.247	482.068	484.273	489.365	482.92
deciduous_timber_m3_ha	inf	inf	inf	inf	inf	inf	inf	inf	inf	inf	inf
mean_dbh_cm	54.355	50.117	54.355	54.355	54.263	54.613	53.832	49.973	56.752	48.77	44.06
pine_mean_dbh_cm	66.035	61.839	66.035	66.035	66.004	66.792	60.984	57.835	63.683	56.946	53.122
spruce_mean_dbh_cm	167.125	155.469	167.125	167.125	166.577	167.881	136.102	132.004	135.442	135.009	129.943
mean_height_m	50.884	46.775	50.884	50.884	50.57	50.855	50.811	47.309	53.049	44.113	39.698
pine_mean_height_m	59.48	55.302	59.48	59.48	58.981	59.441	58.229	55.725	60.294	52.322	49.334
spruce_mean_height_m	156.259	143.707	156.259	156.259	155.15	156.353	130.221	126.402	129.268	128.014	123.053
total_ba_m2_ha	62.188	53.991	59.819	57.122	66.373	57.794	61.602	56.754	59.357	59.825	52.908
pine_ba_m2_ha	79.499	69.991	79.214	78.334	80.131	78.557	79.36	78.205	79.091	75.151	73.518

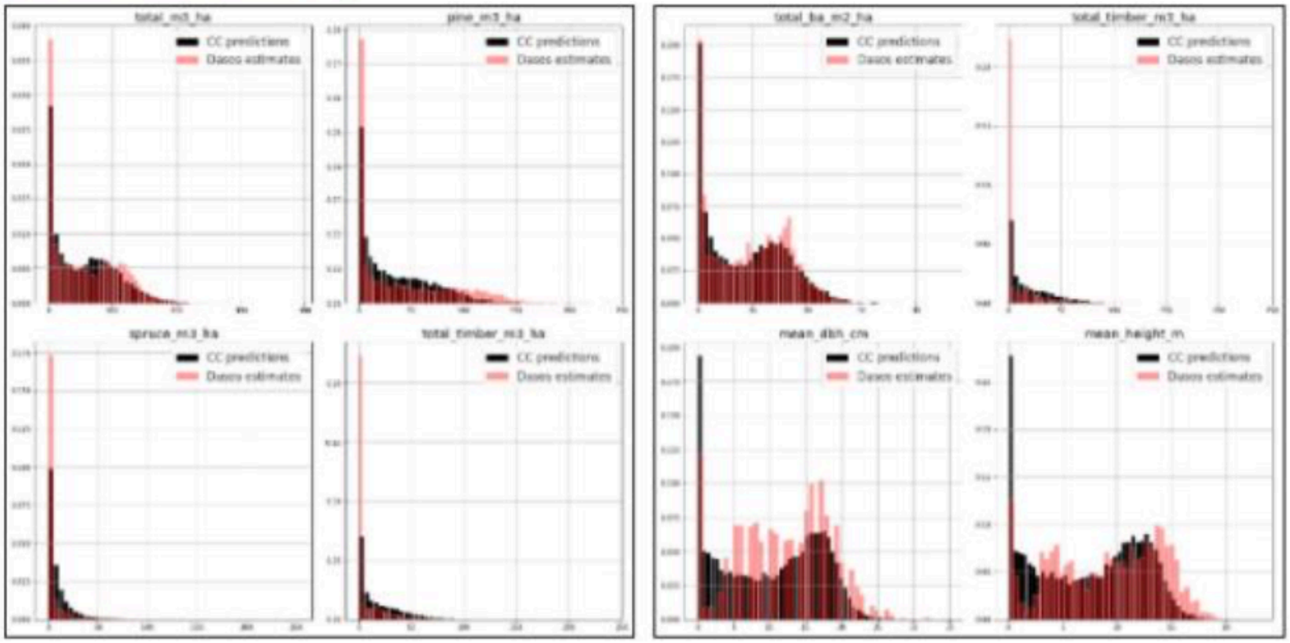
Client Estimate Comparison Report provides comparions of the models' predictions with client estimations. It is used as a sales tool to convince clients that ForestAI's models are better. The reports use Scatter plots and Distrubition plots as shown below.

V103

Predictions CC vs Estimate[redacted]



Distribution Predictions CC vs Estimate [redacted]



Data Modeling Task

You are tasked with designing a data warehouse for implementing these dashboards in a more seamless, systematic, and efficient way. In order to do that, you will first need to **reverse-engineer the existing reports**, then supplement them with other metrics and dimensions based on the expressed business needs. Here are the main subtasks that you need to perform:

- ✓ Identify the metrics that will be visualized in the Dashboards and reports of the BI solution. You can do some research on existing MLOps solutions.
- ✓ Identify the analysis axes that can be used to further analyze these metrics
- ✓ Identify the granularity of your facts (using the multi-dimensional matrix)
- ✓ Create your data modeling schema

Implementation Task

Once the design phase done, you are tasked to implement the Data warehousing/BI solution. According to the expressed business needs, you will need to:

- ✓ Create thematic Datamart(s) containing fact table(s) and dimension tables.

- ✓ Load data into the dimension tables and the fact table(s).
- ✓ Retrieve data from the Datamart(s): creation of a cube + Reporting.

You will need to take into account the following specs:

- ✓ The data loading into the datamart must be done via SSIS
- ✓ The source system should be a relational database that you will need to create and populate with test data
- ✓ Date Dimension's data should be generated using SQL code in a SQL Job on SSIS from a specified start date.
- ✓ At least one dimension in Delta mode: Slowly Changing Dimension (tracking modifications as well as historical data of deleted rows in the original table). Choose the historical attributes wisely.

Data visualization can be done dynamically using an SSAS cube or using Reports/Dashboards (You can create additional reports to those in the specs).

The project needs to be implemented using **Microsoft SQL Server**. Dashboards can be implemented using a platform of your choice.

Delivrables

Deliverables:

- ✓ A brief report describing the conceptual choices, the data model, etc., and explaining the various implementation steps as well as the interpretation of some results regarding the retrieval.
- ✓ Code in a zip file
- ✓ Video Demonstration of the DWH/BI solution: ETL withe execution, Reports and Dashboards (maximum of 10 minutes)

Grading

- ✓ Validity of the Data modeling and pertinence/argumentation of design choices
- ✓ Comprehensiveness: Coverage of all project requirements
- ✓ User-friendliness during data retrieval
- ✓ Ingenuity

The project is done in **groups of three.**