Collections Score Modeling

Consumer Finance Company

Created a score to rank the workflows generated when customers miss their payments thereby helping in reducing the number of defaults

Collection score model for a consumer finance firm

Situation

- Client’s collections team was responsible for reaching out to customers who missed their payments and getting them back on track to successful payment. With the growing customer base, collections team’s time had become a limited resource.
- Partnered with the client to build a scoring model leveraging historical data that would rank the workflows (customers with missed payments) based on probability of default to help the collections team prioritize among customers and reduce defaults

Accordion Value Add

- Analyzed and identified attributes that impact payments/customer defaults such as Customer credit scores, Application & Funding data, Behavioral data, Interaction data, #dishonors, loan application type, employment status of the customer etc.
- Developed iterative models in phases, first for generating the probability of default for a workflow (PD) and second for generating decrease in the probability of default due to a successful call interaction (PD diff * Prob. of answering call)
- Deployed all the models in Azure to run daily to rank the customers with missed payments for the collection team to prioritize
- Developed pipelines to monitor model performance and data drift to run on a weekly basis and raise triggers to identify model retraining

Impact

- 30% Lift in Gini coefficient for PD model compared to the static model clients were using to prioritize customers with missed payments. This indicates that our model separates the default and non default customers 30% better than the static model.
- Customers in the top decile were observed to have a lower propensity to default (45% vs. 20%) when call interaction outcome was included as input to the model

Approach & Methodology

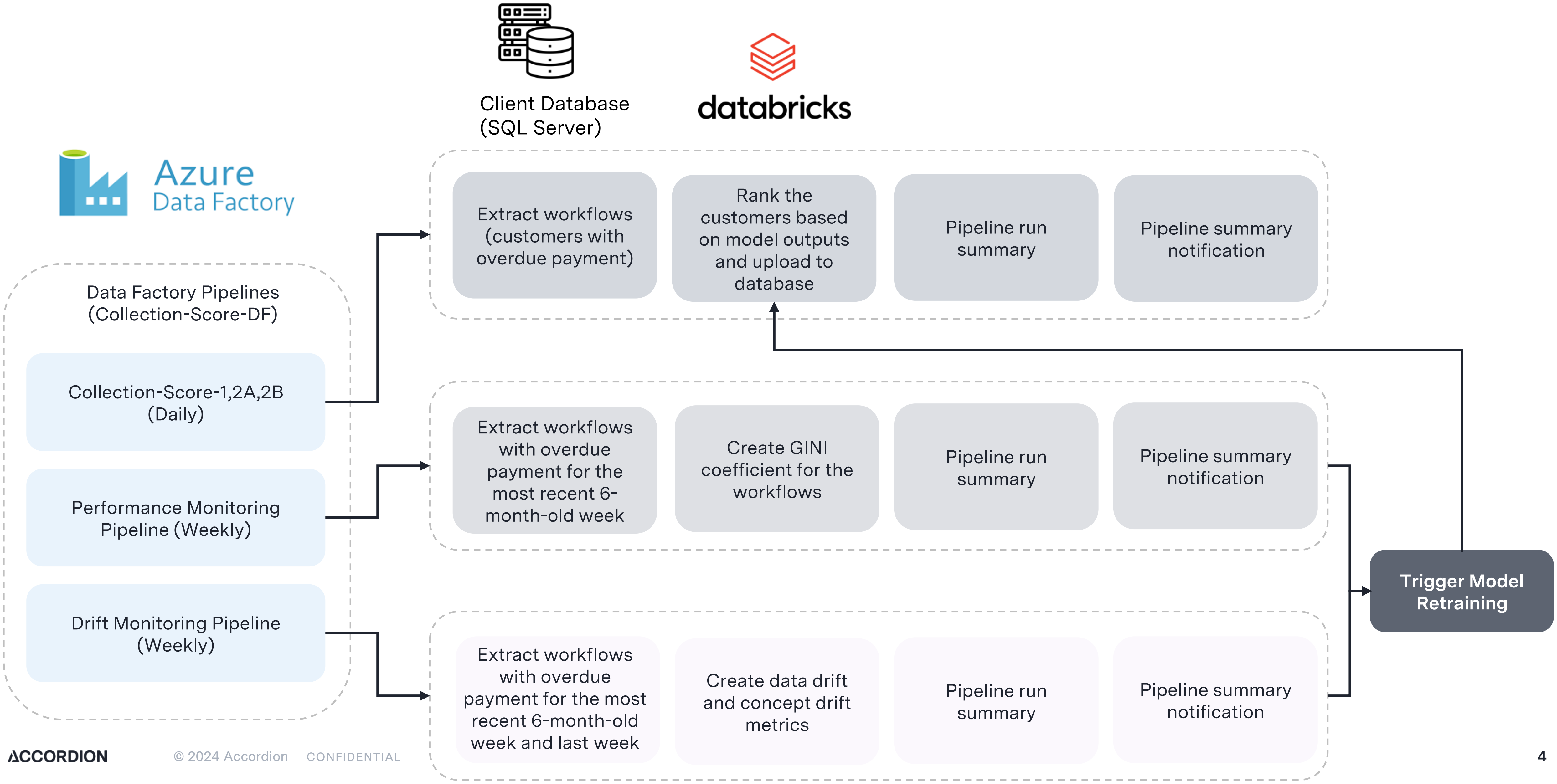
Phase 1 (PD)	Phase 2A (PD Diff)	Phase 2B (PA)
Probability of the customer defaulting in **the next 6 months** when they miss a payment	**Decrease in probability** of customer defaulting if there will be a successful call interaction	Probability of successful call interaction when the call is attempted
Considered **customers data** at the stage of application, funding and their behaviour over the course of loan payments	Considered customers data at the stage of application, funding and their behaviour over the course of loan payments and the outcome of workflow call interaction	Considered previous workflow data, application, funding and behavioural data
Built an **ML (XGBoost) model** with over 53k missed payment events and 61 features resulting in **lift of 30% in Gini Score compared to static model in use**	Built an ML model with over 45k workflows. Ranking of workflows resulted in lowered cumulative default rates, 26% decrease in top 10% ranked workflows	Built a model with 16k workflows which has the best recall of 83% and AUC (Area Under Curve) – 0.66

Deployed models on **Azure pipeline** which runs daily at scheduled time for newly created workflows

Key Terminologies

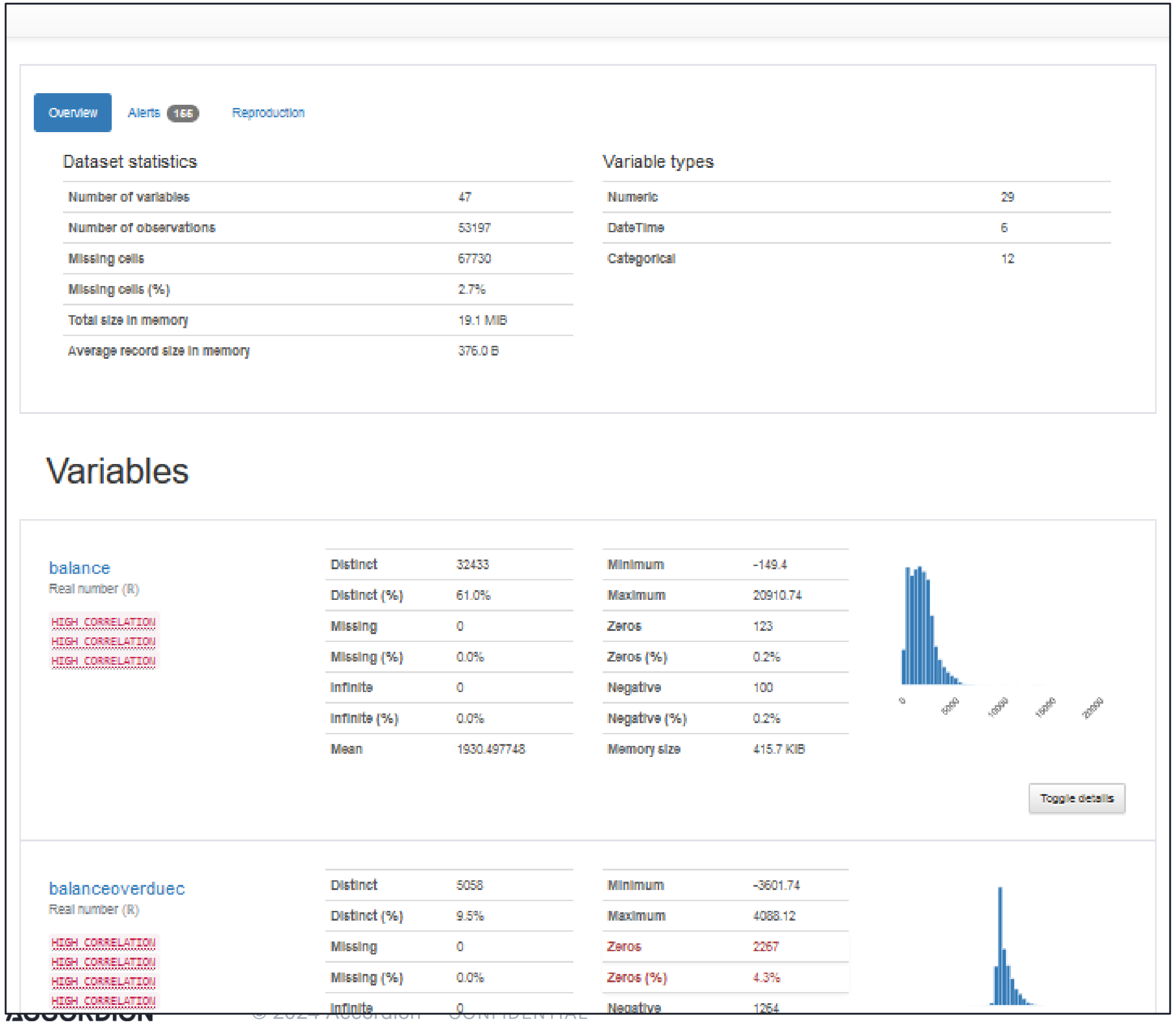
- **Days Past Due (DPD):** No. of days since the last missed payment
- **Probability of Default (PD):** Likelihood of an account going to 90DPD in next 6 months
- **Loss Given Default (LGD):** Losses when the customer defaults
- **Workflow**: Record of a customer with a missed payment. These are generated daily and will be created if a customer with due date on the previous day did not make the payment
- **Expected Credit Loss (ECL):** = Balance x PD x LGD
- **Expected Recoverable Amount (ERA):** = Balance – ECL
- **Ranking is based on impact of call action:** (ERA (call answered) – ERA (call unanswered)) * Probability of answering the call

Collections score architecture



Data reports

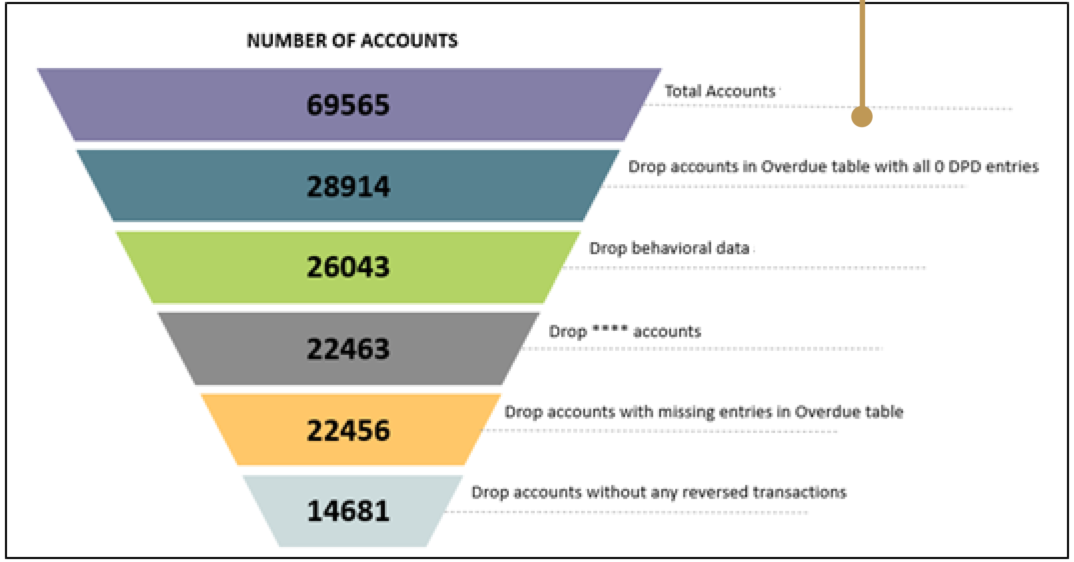
Data Quality Report



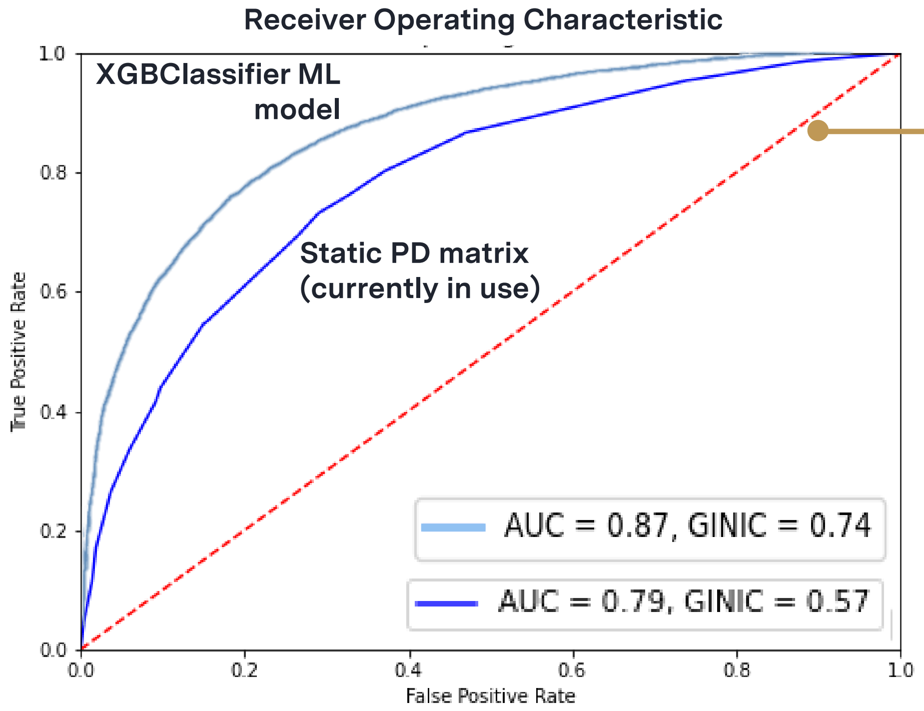
Data Quality reports are generated for each phase at different stages of development to ensure data cleaning and feature engineering

Data funnels are maintained through the development to keep track of data used for training and development

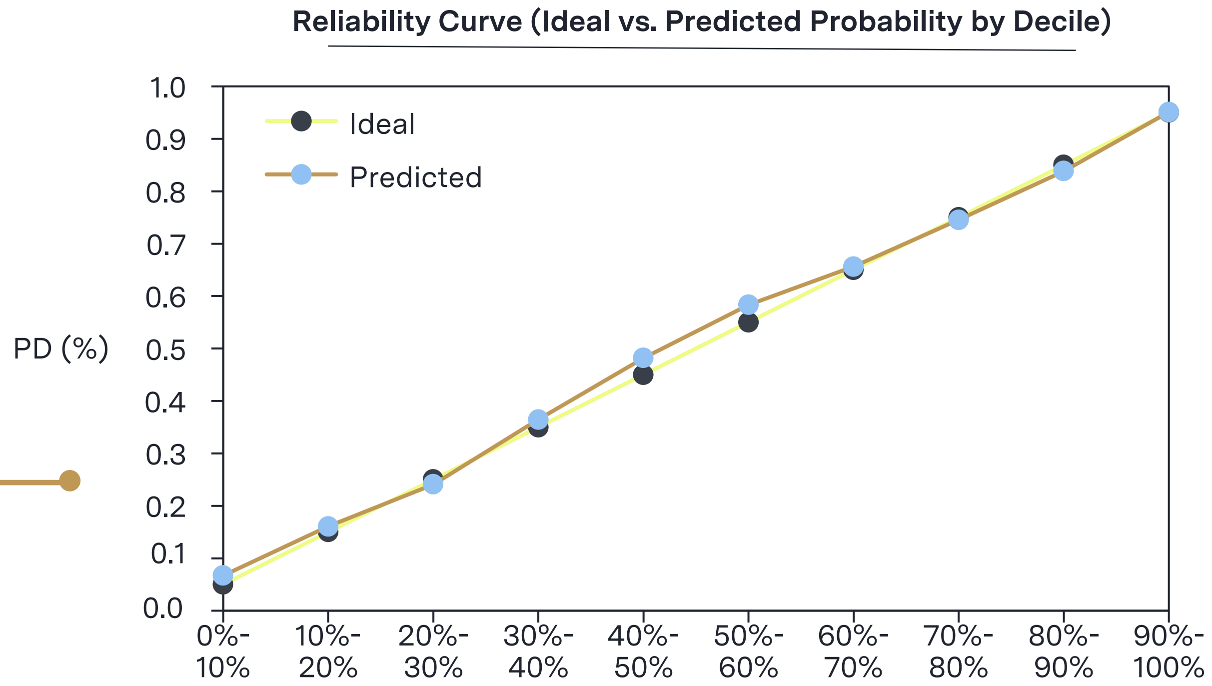
Data Funnel



Phase 1 (PD)

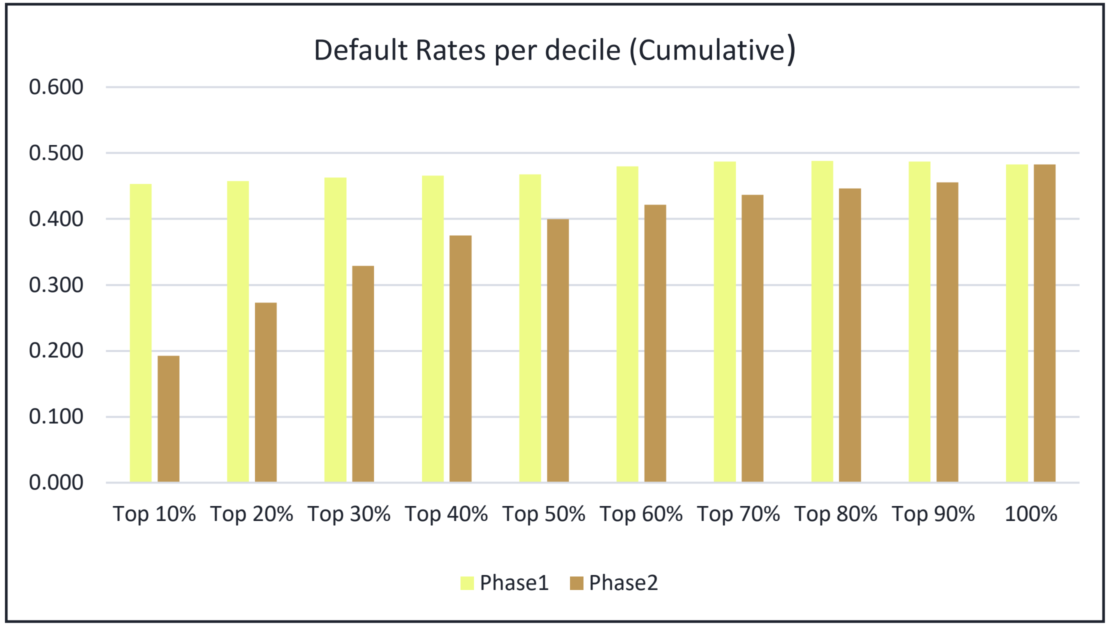


ROC curve for the ML model compared with static model in use. Lift in Gini of 30% . Gini is the measure for how well the model can differentiate default vs non default accounts. This shows that the XGBClassifier model can separate the default customers from non default customers 30% better than the static model.



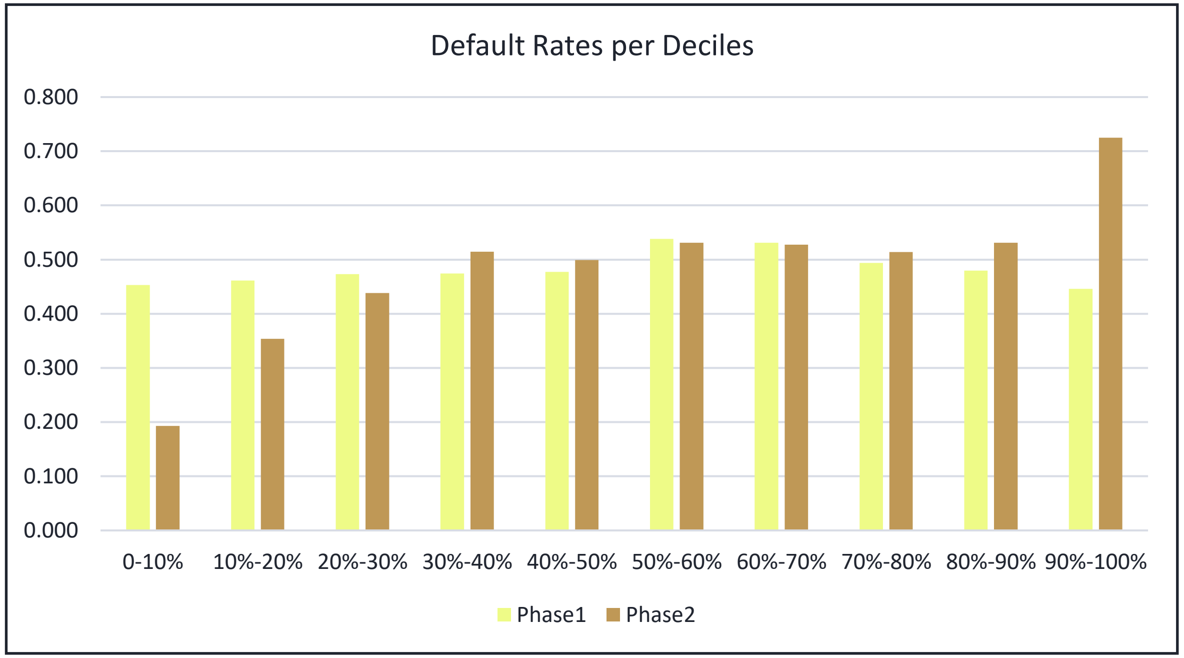
Reliability curve to see how the model is calibrated with respect to ideal distribution

Phase 2A (PD Diff)



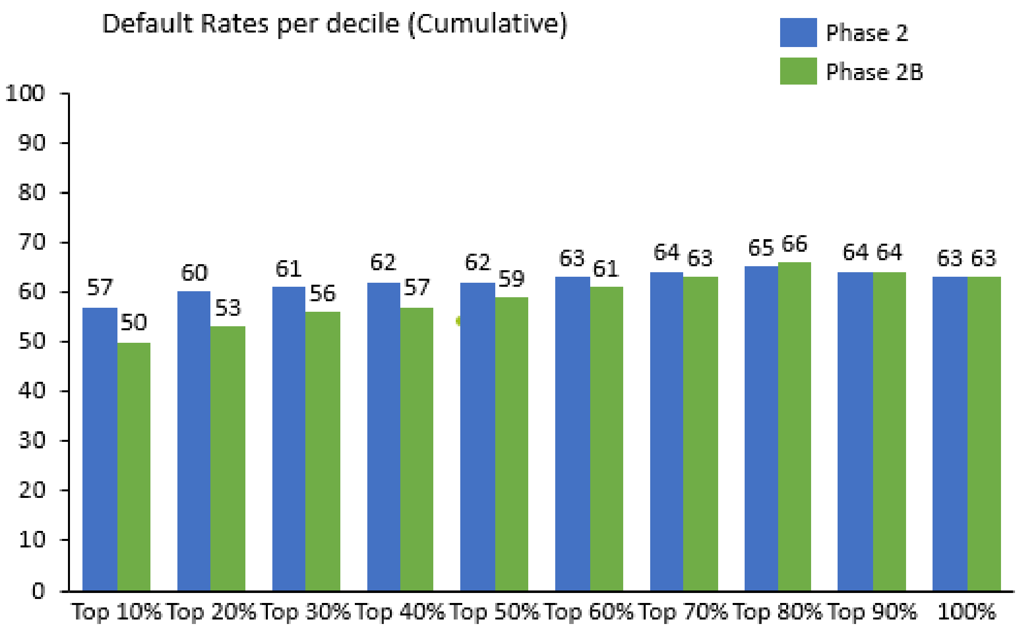
Cumulative default rate is taken for top 10% (based on rank) workflows till 100% where the default rate converges. Evidence of decreased default rate when ranked based on Model outcome

Default rates at different deciles after ranking the workflows. Last decile has high default rate for phase 2 indicating it deprioritizes workflows that do not respond to call action



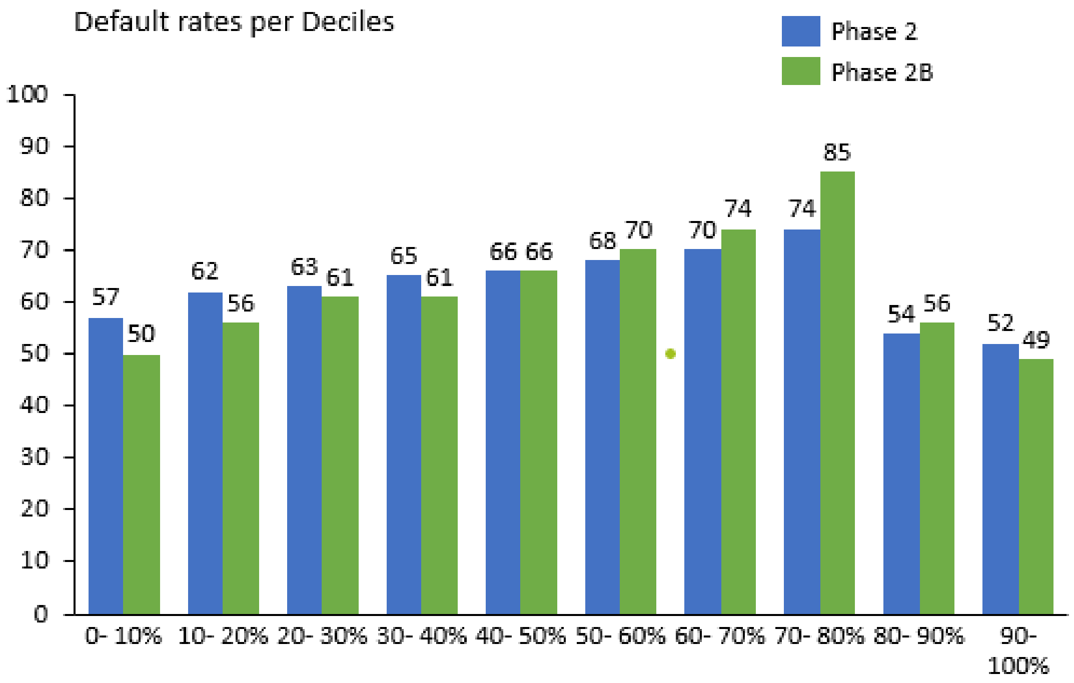
ILLUSTRATIVE

Phase 2B



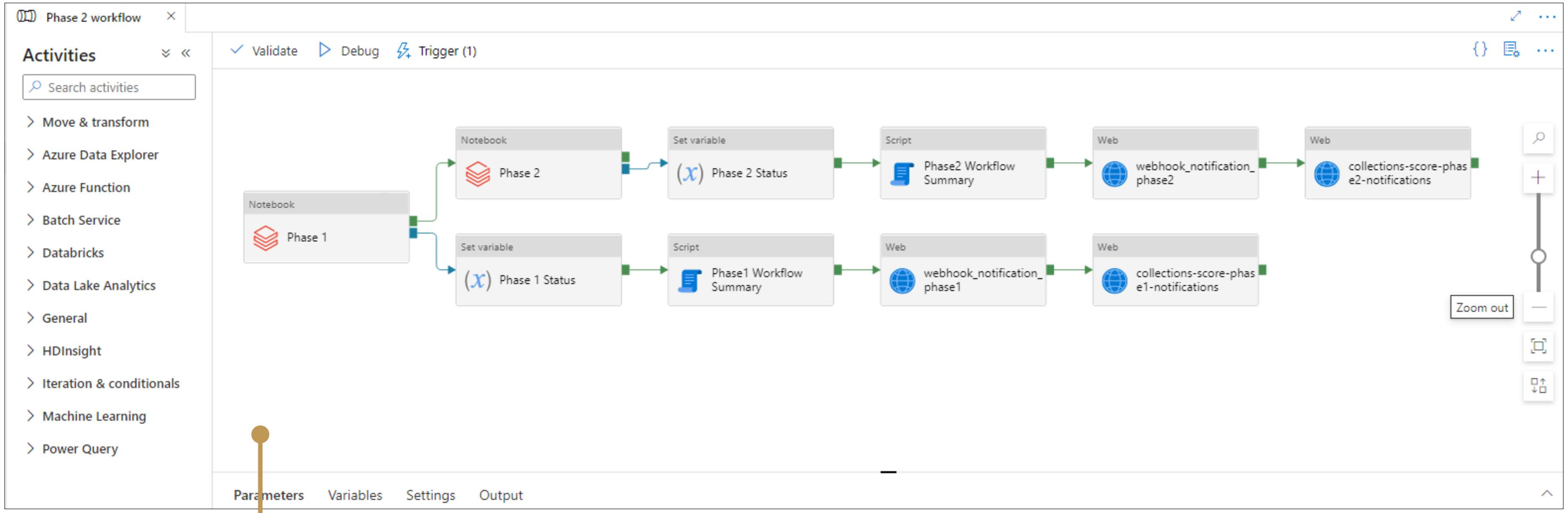
Cumulative default rate is taken for top 10% (based on rank) workflows till 100% where the default rate converges. Evidence of marginal decrease in default rate using phase 2B model

Default rates at different deciles after ranking the workflows. Lower Default rate in last workflows indicative of self healing accounts.



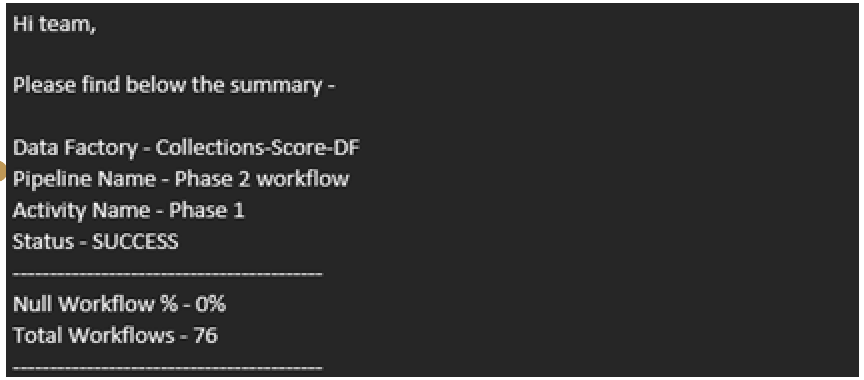
ILLUSTRATIVE

Deployment on data factory



Azure Data Factory pipeline which runs different models and captures the pipeline success or issues

Notification email sent with pipelines execution status


Hi team,

Please find below the summary -

Data Factory - Collections-Score-DF
Pipeline Name - Phase 2 workflow
Activity Name - Phase 1
Status - SUCCESS

Null Workflow % - 0%
Total Workflows - 76
