JUNIOR DATA SCIENTIST - MEXICO

Tech assessment – Nova Edge Customer churn analysis

Marina ATANGANA – 27/06/25

AGENDA

1. Project Introduction & Goal
2. Understanding & Preparing the Data
3. Exploring Customer Behavior (EDA)
4. Our Predictive Modeling Approach
5. Key Findings: What Predicts Churn & Who is High Risk?
6. Business Recommendations
7. Future Work & Improvements
8. Conclusion & Q&A

WHY CHURN MATTERS AND WHAT IS OUR GOAL ?

- Customer churn (customers leaving) is a big challenge for businesses.
- Losing customers means lost revenue and higher acquisition costs.
- **Our Goal**: Use data to understand why customers churn and build a model to predict who will churn.

UNDERSTANDING THE DATA

- We used a dataset with information about customer behavior and characteristics.
- Examples: Transaction volume, activity days, services used, customer notes, segment, region, etc.
- Dataset size: 5000 customers/rows and 11 columns (+1 column for the ID)

[13]
```
# Data preview Display the first 5 rows of the dataset
df.head()
```

	customer_id	monthly_txn_volume	avg_days_active	num_services_used	has_mobile_app	complaints_last_3mo	received_retention_offer	churned	segment	region	industry_type	customer_notes
0	CUST_00000	2872.42	22.0	1	1	1.0	0	0	Mid	CDMX	Healthcare	no contact
1	CUST_00001	1793.36	24.0	4	1	1.0	0	1	Mid	Querétaro	Healthcare	Late Payment
2	CUST_00002	1658.74	26.0	2	1	0.0	0	1	Mid	CDMX	Logistics	No recent activity
3	CUST_00003	1658.76	19.0	4	0	0.0	1	0	Mid	Jalisco	Services	Late Payment
4	CUST_00004	5579.66	22.0	2	0	0.0	0	0	High	CDMX	Logistics	Potential Upsell

Edenred

CLEANING & PREPARING DATA

- **Outlier Handling**: We identified and handled extreme values (outliers) in numerical columns by replacing them with the median, as the median is less affected by extremes.

- **Missing Values:** Some data was missing (e.g., in transaction volume). We filled missing numbers with the median and categorized missing notes.

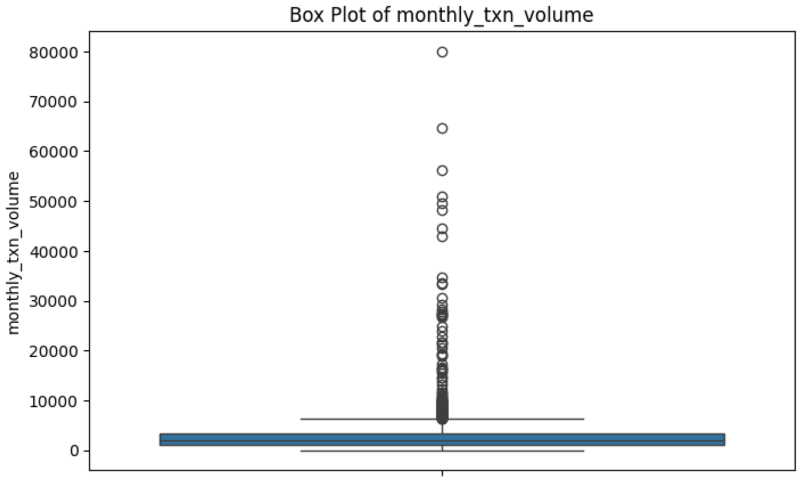
```
# Before the imputation Show percentage of missing values per column
(df.isnull().sum() / len(df)).sort_values(ascending=False) * 100
```

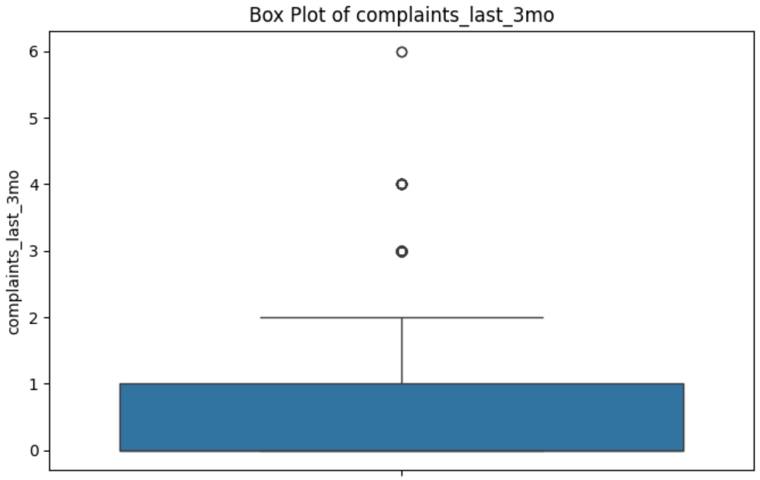
	0
customer_notes	8.58
monthly_txn_volume	5.00
complaints_last_3mo	5.00
avg_days_active	5.00
num_services_used	0.00
customer_id	0.00
received_retention_offer	0.00
has_mobile_app	0.00
churned	0.00
segment	0.00
region	0.00
industry_type	0.00

dtype: float64

	0
customer_notes	8.58
monthly_txn_volume	0.00
avg_days_active	0.00
num_services_used	0.00
customer_id	0.00
has_mobile_app	0.00
complaints_last_3mo	0.00
churned	0.00
received_retention_offer	0.00
segment	0.00
region	0.00
industry_type	0.00
customerNotesCategories	0.00

dtype: float64





CLEANING & PREPARING DATA

- Feature Engineering: We created a new feature from customer notes to make them useful for the model (e.g., categorizing notes like 'Complaint', 'Engagement Issue').

- Preprocessing: Transformed data for the models (e.g., scaling numbers, encoding categories).

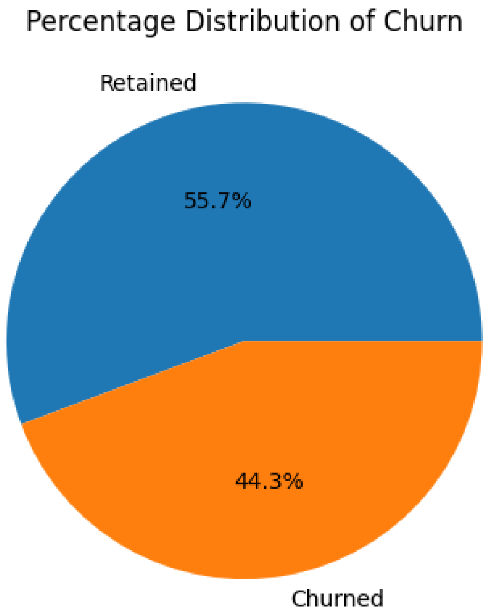
```
       customer_notes     customerNotesCategories
0          no contact           Engagement Status
1        Late Payment   Financial/Payment Related
2  No recent activity           Engagement Status
3        Late Payment   Financial/Payment Related
4     Potential Upsell  Sales/Upsell Opportunities
```

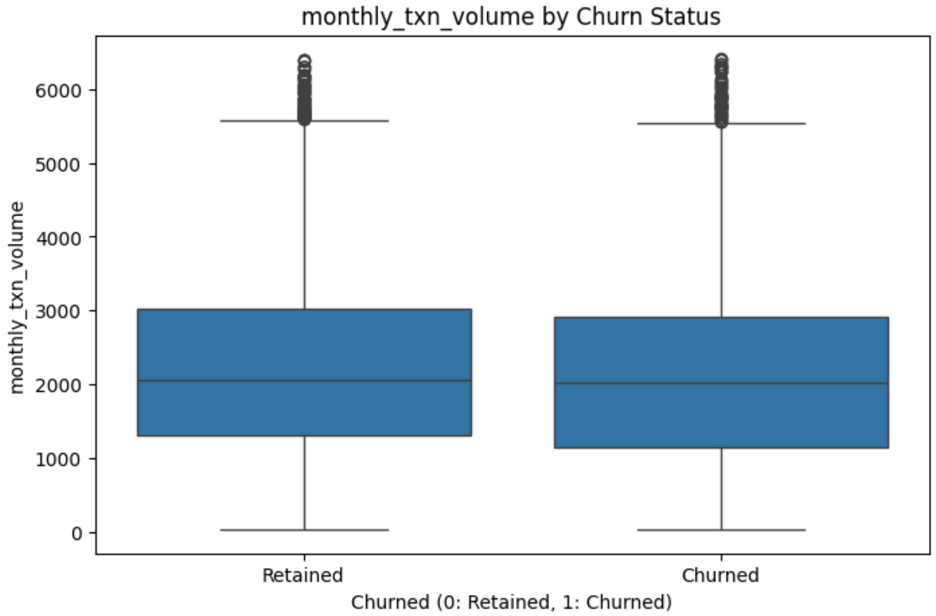
received_retention_offer	segment_High	segment_Low	segment_Mid	region_Baja California
0.0	0.0	0.0	1.0	0.0
0.0	0.0	0.0	1.0	0.0
0.0	0.0	0.0	1.0	0.0
0.0	0.0	0.0	1.0	0.0
0.0	1.0	0.0	0.0	0.0

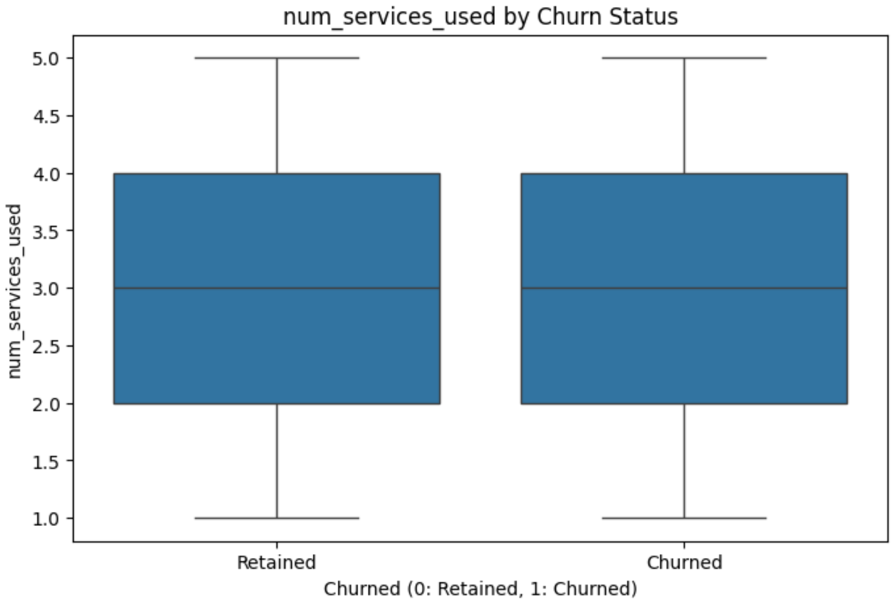
DATA EXPLORATION

- We looked at **how many customers churned vs. stayed.**
- We examined how different features looked for churned vs. retained customers.

- **Initial look at individual features didn't show huge differences on their own.**







WHY PREDICTIVE MODELING?

- Simple data views didn't show strong individual predictors.
- **Churn is likely caused by *combinations* of factors.**
- **We need models to find these complex patterns and predict churn likelihood.**

ASSESSING OUR PREDICTIVE POWER (MODEL PERFORMANCE)

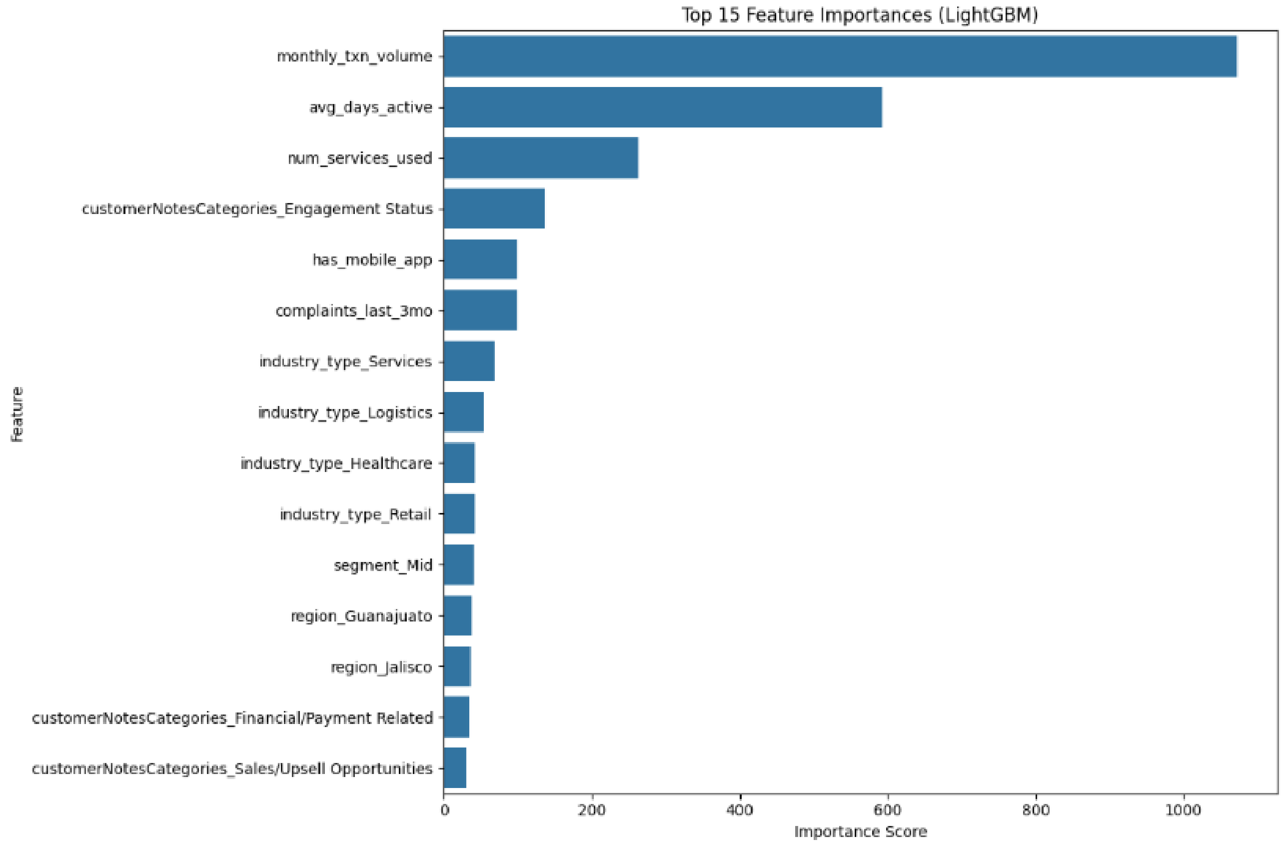
- **Rigorous Testing:** Data split into Training (75%) for learning, and unseen Testing (25%) for unbiased evaluation.
  - **Key Metrics:** Evaluated models using:
  - **Recall (Churners Found):** How many actual churners did we correctly identify? (Crucial for early intervention)
  - **Precision (Accurate Churn Flags):** Of those flagged as churners, how many *churned*? (Important to avoid wasted efforts)
  - **F1-Score:** A balanced measure of both Recall and Precision.
  - *Accuracy and Confusion Matrix also considered.*
- **Overall Challenge:** Accurately predicting churn with this dataset proved challenging; no model achieved very high F1-scores (e.g., above 0.75).
- **Model Strengths:**
  - **Neural Network:** Best at **Recall** (catching more actual churners) but also had more "false alarms."
  - **Random Forest & LightGBM:** Offered a better **balance** of Recall and Precision, making them more practical for targeted actions.

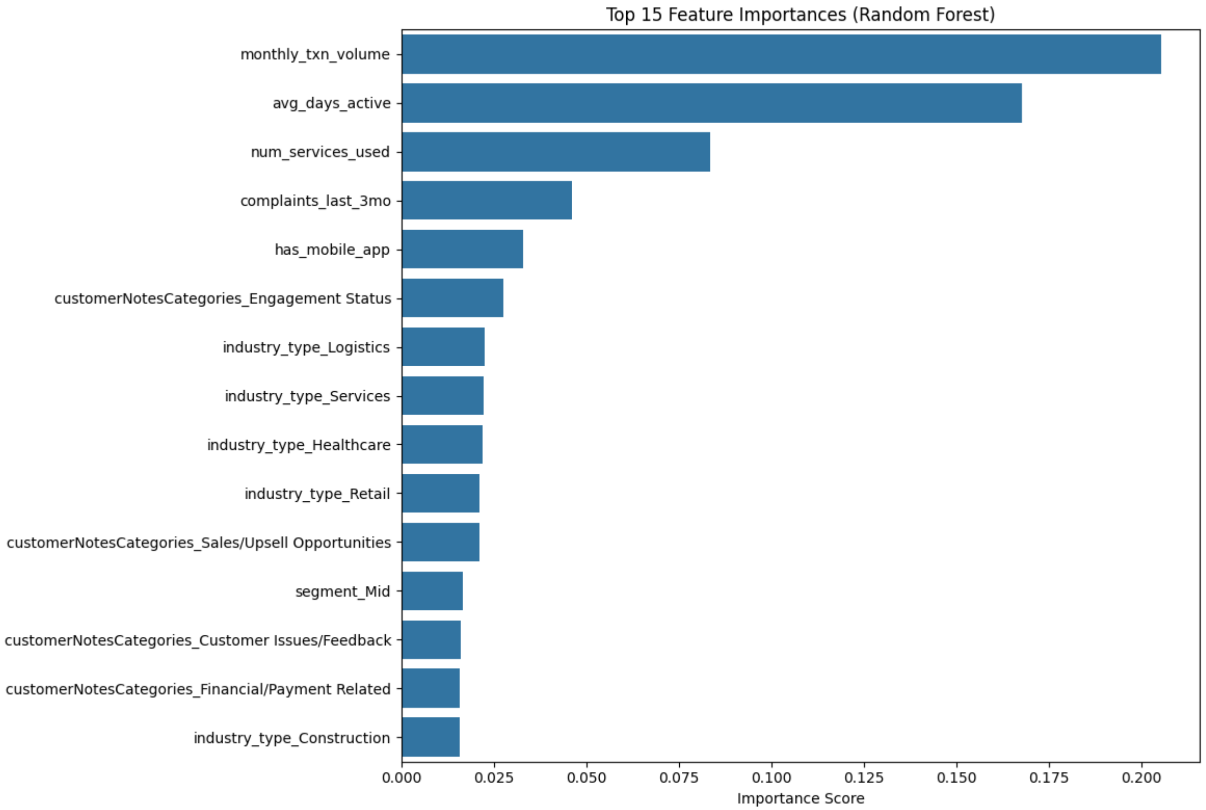
	Model	F1-Score (Churn)	Recall (Churn)	Precision (Churn)	False Positives
0	Logistic Regression	0.296954	0.211191	0.500000	117
1	Random Forest	0.416490	0.355596	0.502551	195
2	LightGBM	0.403361	0.346570	0.482412	206
3	Neural Network	0.451253	0.438628	0.464627	280

Edenred

WHAT PREDICTS CHURN?

- **Insights from Top Models:** Analysis based on the robust Random Forest and LightGBM models, which provide clear insights into feature importance.
- **Dominant Behavioral Predictors:**
  - **Low Usage Volume:** Customers with significantly lower monthly transaction volumes are the strongest indicator of churn risk.
  - **Activity Levels:** Infrequent engagement and fewer active days on the service are highly predictive.
  - **Services Utilized:** A lower number of different services actively used also signals increased risk.
  - **Critical Customer Interaction Signals:**
  - **Recent Complaints:** The presence of recent customer complaints serves as a strong, immediate warning sign.
  - **Specific Customer Note Categories:** What is recorded in customer notes, particularly concerning 'Engagement Status' (e.g., indicating inactivity or disinterest) and 'Financial/Payment Issues', are highly influential churn predictors.
- **Less Impactful Factors (on their own):** General demographic or static information like Industry, Segment, or Region showed less direct predictive power when analyzed in isolation.





IDENTIFYING OUR MOST CHURNABLE CUSTOMERS

•**Individual Churn Probability:** Our models assign a churn probability score to every customer.
•**Actionable High-Risk Lists:** We can now identify and prioritize customers with the highest predicted risk (e.g., the top 10-20% most likely to churn).
•**Consistent Profile:** High-risk customers consistently exhibit the key churn drivers:
- **Subdued Activity:** Low usage volume and infrequent activity.
- **Expressed Dissatisfaction:** Presence of recent complaints.
- **Documented Concerns:** Customer notes indicating low engagement or financial challenges.

•**Cross-Model Validation:** While different models might highlight slightly different customers, there's significant overlap among the top-risk individuals, especially those with multiple warning signs.

Top 20 Churn Risk Customers (based on average predicted probability from RF, LGBM, NN):

	customer_id	average_churn_probability	churn_probability_rf	churn_probability_lgbm	churn_probability_nn
1861	CUST_01861	0.806659	0.910	0.907343	0.602635
725	CUST_00725	0.797225	0.950	0.864956	0.576718
2397	CUST_02397	0.796225	0.925	0.867762	0.595913
76	CUST_00076	0.787710	0.870	0.861192	0.631939
4762	CUST_04762	0.781201	0.880	0.833599	0.630005
4811	CUST_04811	0.777833	0.900	0.836401	0.597098
2518	CUST_02518	0.776511	0.870	0.810896	0.648638
3737	CUST_03737	0.771708	0.910	0.824669	0.580455
4304	CUST_04304	0.771672	0.920	0.815788	0.579229
3023	CUST_03023	0.770679	0.870	0.794342	0.647696
3514	CUST_03514	0.770060	0.920	0.767161	0.623020
222	CUST_00222	0.766319	0.920	0.858272	0.520685
3476	CUST_03476	0.765546	0.880	0.829309	0.587330
1315	CUST_01315	0.760888	0.880	0.847515	0.555148
1283	CUST_01283	0.760736	0.860	0.792612	0.629594
1088	CUST_01088	0.758619	0.890	0.720457	0.665398
2166	CUST_02166	0.757706	0.880	0.857356	0.535762
2179	CUST_02179	0.756067	0.860	0.836970	0.571232
2481	CUST_02481	0.755803	0.890	0.796299	0.581111
2449	CUST_02449	0.752428	0.800	0.878805	0.578479

TURNING INSIGHT INTO ACTION

- **Target Low Users & Inactive:** Proactively engage customers with low transaction volume and activity. (eg: offer incentives to boost usage.)
- **Address Service Gaps:** Help customers use more services; highlight benefits they might be missing.
- **Improve Complaint Handling:** Make resolving issues for customers with complaints a top priority; follow up quickly.
- **Act on Customer Notes:** Use the categories from notes (engagement, financial, issues) to personalize outreach and solve specific problems.
- **Build Alert System:** Implement an automated system to notify Sales/CS teams immediately when a customer is flagged as high risk.

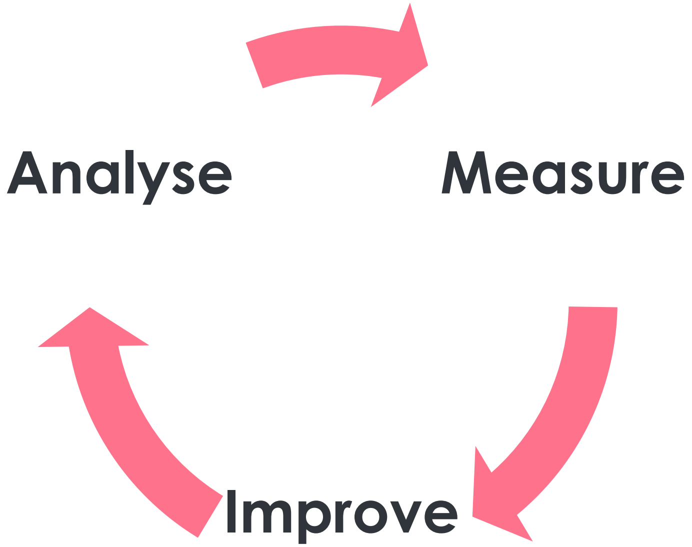
WHAT'S NEXT?

- **Implement & Measure:** Put recommendations into practice and measure impact (A/B testing).

- **Better Data:** Get more granular transaction/interaction data, time-series data to capture trends.

- **Model Enhancement:** Tune models further, explore other advanced techniques or model ensembles.

Cluster Analysis: Explore customer groups using clustering for new insights into churn causes.



Edenred

CONCLUSION & KEY TAKEAWAYS

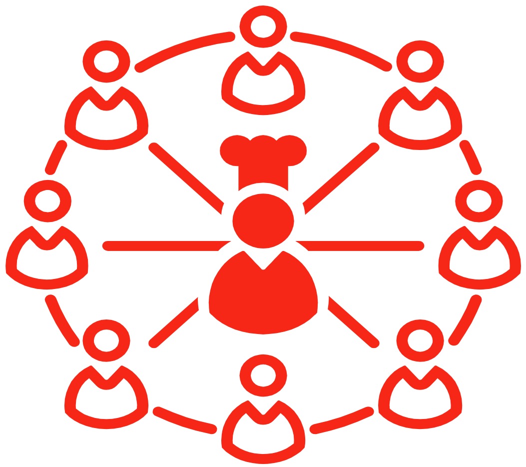
- We used data and ML models to predict churn.

- Key churn drivers identified: **low usage, complaints, engagement/financial notes**

- We can identify **high-risk customers** to focus on.

- **Targeted actions** based on these insights are recommended to improve retention.

- Success requires implementing actions and **measuring their impact**.



Edenred

Q&A