

Data Science & AI - Case Assignment

Suraj Karakulath, August 2025



TOC

Brief	Model training & comparison	Email generation prompt
Part I - Churn Prediction	Results and Insights	Email generation prompt II
Data exploration	Part II - Email generation	Adherence to brand guidelines
Feature engineering	Prompt engineering	Scaling across journeys



Brief

A new client of DEPT faces challenges in retaining customers, leading to increased churn rates and lost revenue. DEPT wants to leverage ML and AI to identify customers at risk of leaving and take proactive measures to retain them.

This project is a proof of concept to demonstrate how our AI application can

1. Predict customer churn based on various attributes
2. Generate email content for retention send-outs that adhere to brand guidelines, using LLMs in order to enable targeted retention strategies and enhance customer satisfaction.

A short horizontal bar with a teal segment on the left and an orange segment on the right.

Part I - Churn Prediction



Process workflow

Data exploration

- Feature distributions
- Outliers
- Null values
- Target variable imbalance

Feature engineering

- Imputing null values
- One-hot encoding categorical variables

Model training (Baseline model)

- Start with baseline model results
- Decide on key metric

Model comparison and evaluation

- Compare key metric with other models
- Find best parameters for best model using (GridSearchCV)

Results and Insights

- Model interpretability
- Most important factors influencing churn



Data exploration

1

7043 entries with as many unique customerIDs. That means there are no duplicate customerIDs.

2

3 numerical features - Tenure, MonthlyCharges, and TotalCharges (this was object type - so had to convert to numeric)

3

16 categorical features - 6 of them binary and 10 of them with 3 or 4 categories.

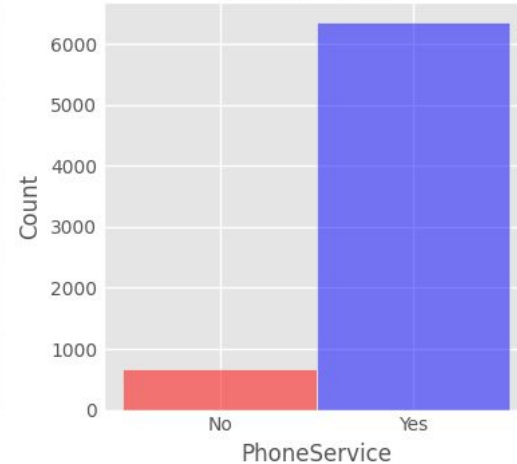
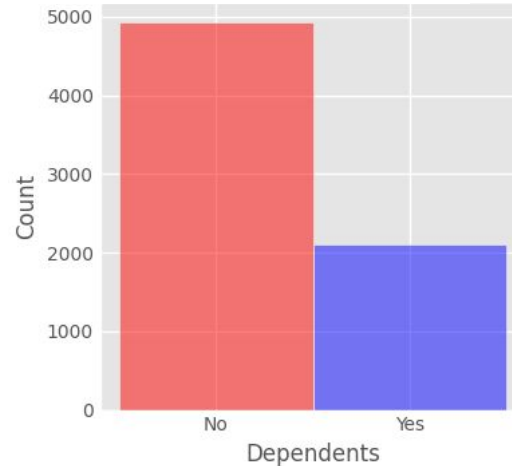
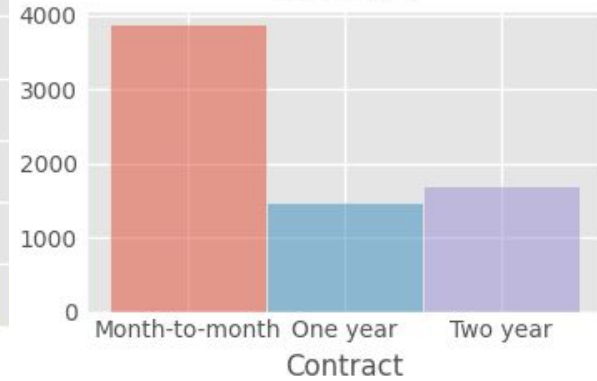
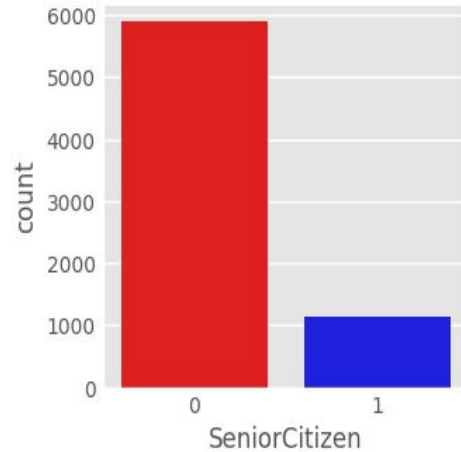
4

Target variable Churn is binary - Yes or No

Data exploration

- Fewer Senior Citizens than non-Senior Citizens.
- More customers with no dependents
- More customers with Phone service
- More customers with month-to-month contract

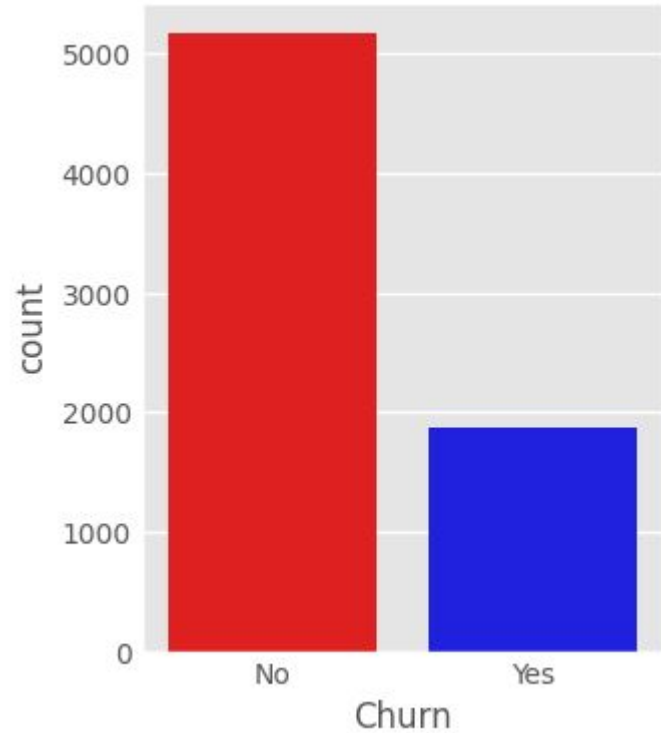
These are only good to know at this stage
(to be validated in prediction later).





Data exploration

- Key observation is that the Churn data is somewhat **imbalanced**.
- So we use **class weights** in model training.





Feature engineering

- There were 11 NaN values for TotalCharges
- On further examination, these were **for tenure = 0** (brand new customers).
- So instead of removing these rows, we set **TotalCharges here = 0** (billing cycle not started yet)

customerID	tenure	MonthlyCharges	TotalCharges
4472-LVYGI	0	52.55	NaN
3115-CZMZD	0	20.25	NaN
5709-LVOEQ	0	80.85	NaN
4367-NUYAO	0	25.75	NaN
1371-DWPAZ	0	56.05	NaN
7644-OMVMY	0	19.85	NaN
3213-VVOLG	0	25.35	NaN
2520-SGTTA	0	20.00	NaN
2923-ARZLG	0	19.70	NaN
4075-WKNIU	0	73.35	NaN
2775-SEFEE	0	61.90	NaN



Model training (Baseline model)

Confusion Matrix:
[[748 287]
[80 294]]

- Start with simple **logistic regression** as baseline
- Train-test split of **80-20**
- **Recall of 79%** is good (we don't want to miss churners)
- There will be **false alarms** (FP = 287) so you'd waste retention efforts on some loyal customers.
- In real life, **this trade-off is OK** - better to email a few loyal customers than miss churners.

Classification	Report:		
	precision	recall	f1-score
0	0.90	0.72	0.80
1	0.51	0.79	0.62
accuracy			0.74
macro avg	0.70	0.75	0.71
weighted avg	0.80	0.74	0.75

ROC AUC Score: 0.8419101500942933



Model comparison

- Main metric to compare is ROC-AUC Score
- Logistic regression, CatBoost and Neural Net and are good candidates for the best models.
- After **hyperparameter** tuning (GridSearch) on these 3, **CatBoost** was the best

	Model	ROC_AUC
0	Logistic Regression	0.841910
1	CatBoost	0.839545
2	Neural Net	0.830866
3	Random Forest	0.824086
4	Naive Bayes	0.823909
5	XGBoost	0.821065
6	kNN	0.754508
7	Decision Tree	0.658082

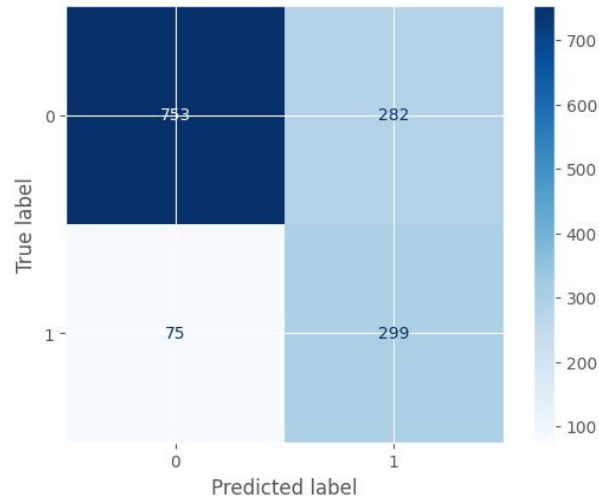


Best model - CatBoost

Using best parameters from parameter tuning for CatBoost:

- Improved recall: 80%
- Fewer false alarms (FP = 282)
- Better ROC AUC Score: 0.84

Better predictive capability than baseline



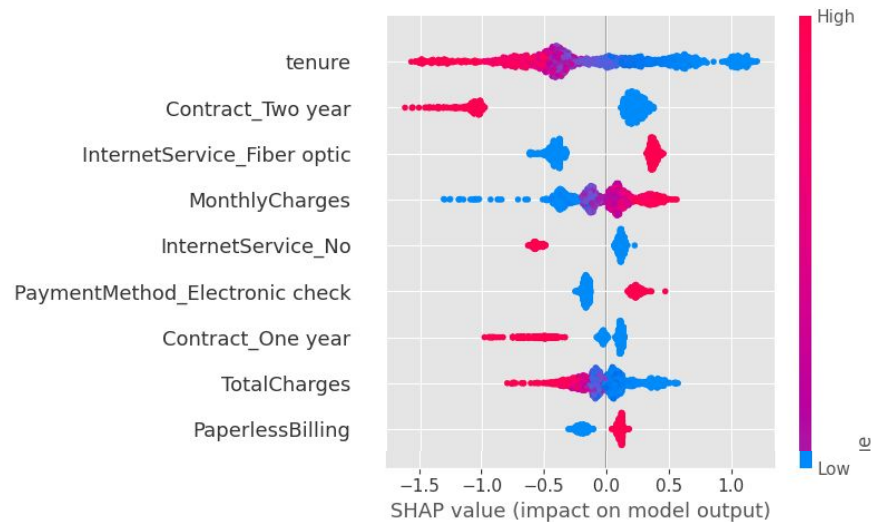
	Classification Report:			
	precision	recall	f1-score	support
0	0.91	0.73	0.81	1035
1	0.51	0.80	0.63	374
accuracy			0.75	1409
macro avg	0.71	0.76	0.72	1409
weighted avg	0.80	0.75	0.76	1409

ROC AUC Score: 0.8472267431346715

Results and Insights

SHAP explainer for model interpretability to find most important features

1. **Tenure:** longer tenure (or more loyal customers) = lower churn risk, shorter tenure or newer customers = higher churn risk.
2. **Contract_Two year** is next: Similar to Tenure. Also clear separation = extremely stable compared to other contracts
3. **InternetService_Fiber optic** - Customers with Fiber optic service more likely to churn compared to the baseline (DSL). Also more stable - clear binary signal for churn.
4. **Monthly charges** - Customers paying more more likely to churn, but not as binary as contract type





Results and Insights

- We can predict if a customer will churn or not (binary).
- We can also bucket customers based on a churn risk probability.

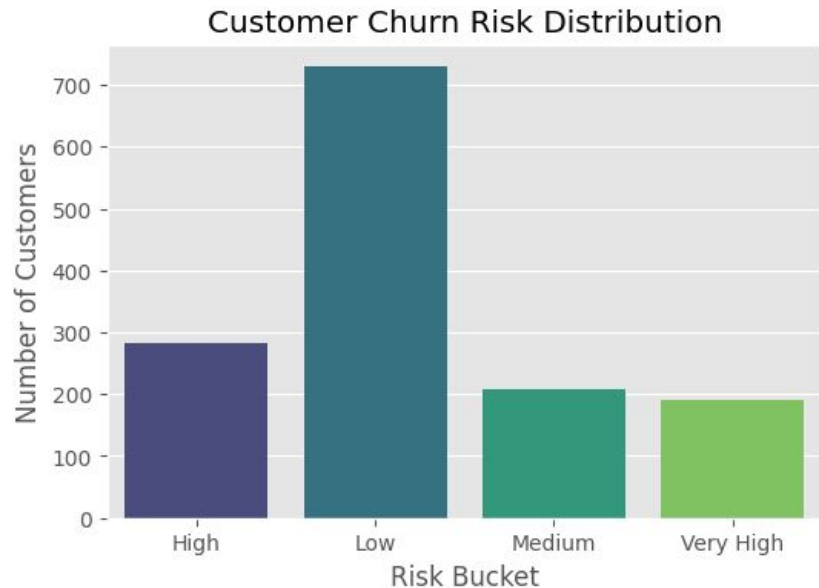
	customerID	Churn_Pred	Churn_Prob	Risk_Bucket
0	4376-KFVRS	0	0.060229	Low
1	2754-SDJRD	1	0.896378	Very High
2	9917-KWRBE	0	0.155458	Low
3	0365-GXEZS	1	0.544443	Medium
4	9385-NXKDA	0	0.039584	Low
...
1404	5204-HMGYF	0	0.109108	Low
1405	9950-MTGYX	0	0.034748	Low
1406	3675-EQOZA	0	0.293385	Low
1407	3646-ITDGM	0	0.035757	Low
1408	3913-FCU UW	0	0.022303	Low



Results and Insights

Based on the risk bucket, we can personalise retention email strategy. For example:

- **Very High:** retention offer + friction-reduction CTA
- **High:** soft offer / value reinforcement
- **Medium:** tips/feature education
- **Low:** loyalty/upsell



A short horizontal bar with a teal segment on the left and an orange segment on the right.

Part II - Email generation



Prompt engineering

1

Adopt a **persona** (“You are an expert copywriter for Vodafone...”)

2

Be **specific** about the **goal** (“...write a customer retention email for users at risk of churning...”)

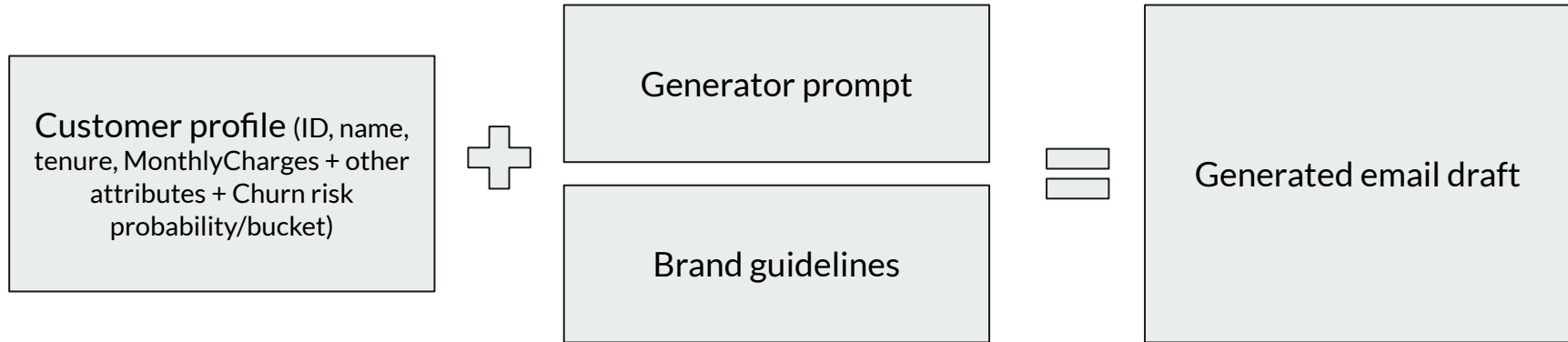
3

Provide as much **context** as necessary (“...be friendly, clear and professional...reference their current service...include a compelling offer for them to stay...”)

4

Specify **constraints** (“...adhering to brand and tone of voice guidelines...” etc.)

Example workflow (Simple approach)



Example prompt (Simple approach)



You are an expert copywriter for a telecommunications company, Vodafone. Your task is to write a customer retention email for a user who has been flagged as high-risk for churning.

The email should be friendly, clear, and professional. It should convey trust, reliability, and appreciation for the customer. To achieve this, the email must reference the services they currently have (e.g., phone, internet, streaming services), remind them of the benefits they receive and include a clear and compelling offer to encourage them to stay (a special discount, an upgrade, or a trial of a new service etc.)

Follow the tone of voice and email structure specified in the Brand Guidelines document. Remember to personalize using the customer's name and tailor the content to their preferences and usage history.

Please generate the email content now.

Example output



Subject: A special thank you for being with Vodafone, Wilhelm!

Hi Wilhelm,

Thank you for being a valued part of the Vodafone family for the past 23 months. We're so grateful to have you with us and we wanted to reach out to you personally to show our appreciation.

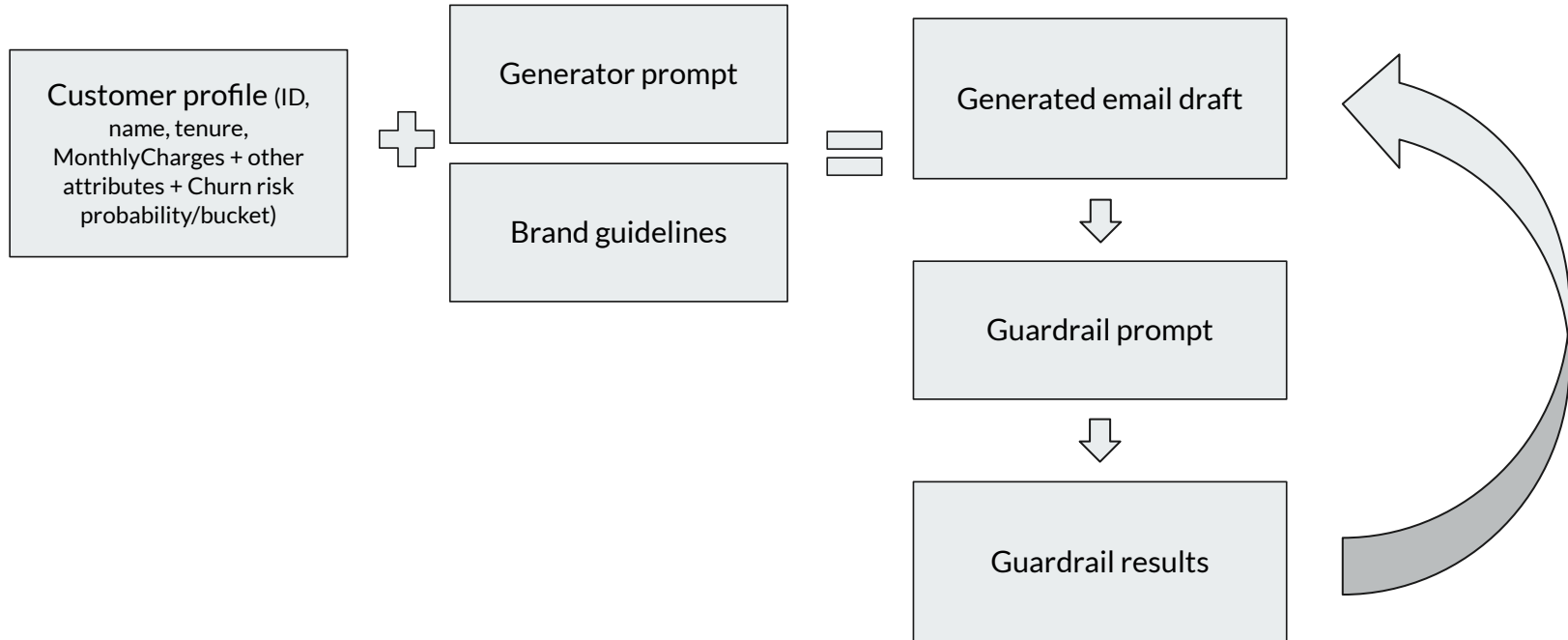
We understand that you've been with us on a month-to-month plan, and we know how important it is for you to have a service that you can rely on without breaking the bank. So, we've prepared a special, exclusive offer just for you to thank you for your loyalty:

- **A 10% discount** on your monthly bill for the next 12 months. That's a total savings of over £35!
- **A free upgrade** to our premium streaming service for three months, giving you access to all the latest movies and TV shows at no extra cost.

To take advantage of this special offer, simply reply to this email or click the button below to claim your rewards. We're committed to providing you with the best possible service, and we hope you'll continue to be part of the Vodafone family for many years to come.

*Warmly,
The Vodafone Team*

Example workflow (2-layer approach)



Example prompts (2-layer approach)

Generator Prompt

You are an email copywriter for Vodafone.

Write a customer retention email that follows Vodafone's Tone of Voice:

Tone attributes:

- *Friendly and approachable: warm, conversational, no jargon.*
- *Clear and concise: short sentences, easy to read.*
- *Positive and reassuring: highlight benefits, empathize with concerns.*
- *Professional and trustworthy: respectful, accurate, reliable.*

Email structure:

1. *Subject line: Friendly, enticing, relevant (max 7 words).*
2. *Greeting: Warm and personalized.*
3. *Introduction: Briefly explain the purpose.*
4. *Body: Highlight offers or updates, use bullet points if needed.*
5. *Call to Action: Clear, compelling.*
6. *Closing: Warm and appreciative.*
7. *Signature: Friendly and professional.*

Output format:

- *2 subject line options*
- *Full email (Greeting → Body → CTA → Closing → Signature)*

Guardrail Prompt

Check if the following email follows Vodafone's Tone of Voice.

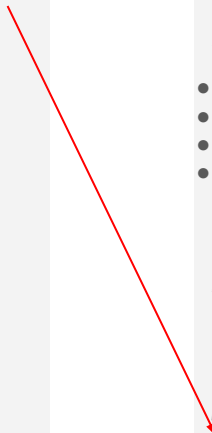
Guidelines:

- *Friendly and approachable: warm, conversational, no jargon.*
- *Clear and concise: short sentences, bullet points if helpful.*
- *Positive and reassuring: highlight benefits, show empathy.*
- *Professional and trustworthy: respectful, accurate.*

For each attribute, say PASS or FAIL and suggest improvements if needed.

Email to check:

[Generated email from Step 1 here]





Ensuring adherence to brand guidelines

1. Manual editing - human in the loop
2. Specify critical points in brand guidelines as constraints
3. Layer 2 for validation with a JSON output that is machine-readable
4. Review Layer 2 output and refine email/prompt to repeat from step 1



Scaling across journeys + other use cases

- **Journeys:** Onboarding nudge (tenure 0–2) or mid-contract/pre-renewal
- **Funnel stages:** Awareness, engagement, retention, loyalty
- **Segments:** High spenders vs. low spenders, young vs. senior, DSL vs. Fiber optic etc.

A similar framework can be used to generate onboarding emails, upsell offers, loyalty rewards, or win-back messages. We will just need to swap the Goal in the prompt accordingly. E.g. “*Welcome a new customer and help them get started.*” instead of encouraging retention.

- Finally, run **A/B tests** and refine prompts, offers etc. to optimise performance over time.



Thank you.

