Assessment

Part 1

1a Task: De-dupe students_activities table using provided logic

SQL code: [Part 1a - clean_students_activities.sql]

```
SELECT id,
       student_id,
       account_id,
       resource_type,
       score,
       quiz_id,
       etl_last_updated_ts
FROM   (SELECT *,
               Row_number()
                 OVER (
                   partition BY id
                   ORDER BY etl_last_updated_ts, account_id) AS row_number
        FROM   students_activities) AS ordered_data
WHERE  ordered_data.row_number = 1
```

1b Task: Generate summary statistics using cleaned students_activities dataset and grouping by resource type[1]

SQL code: [Part 1b - aggregate_students_activities.sql]

```
WITH clean_students_activities
     AS (SELECT id,
                student_id,
                account_id,
                resource_type,
                score,
                quiz_id,
                etl_last_updated_ts
         FROM   (SELECT *,
                        Row_number()
                          OVER (
                            partition BY id
                            ORDER BY etl_last_updated_ts, account_id) AS
                        row_number
                 FROM   students_activities) AS ordered_data
         WHERE  ordered_data.row_number = 1)

SELECT resource_type,
       Count(DISTINCT student_id) AS num_students,
       Round(Avg(score), 2)       AS mean_score,
       Min(etl_last_updated_ts)   AS min_etl_update,
       Max(etl_last_updated_ts)   AS max_etl_update
FROM   clean_students_activities
GROUP  BY resource_type
```

[1] In the sample dataset, all *score* values where *resource_type* = “movie” are blank. If this was the case for the entire dataset I would clarify whether this was expected. With missing scores for one resource type, there would be limited value in this comparison.

Part 2:

Task: Explain how you would identify the most important data points to predict churn and retention

Project prerequisites:

- Define and agree upon terminology and calculations for outcome variable (churn, retention)
- Define and agree upon calculation for any derived input variables (e.g. usage stats)
- Access to data sources, data dictionaries, and associated documentation
- Context on available data sources (historical knowledge, data quality)
- Identified contact(s) for any clarifying questions about data

Statistical / Machine Learning Methods:

1. Generate descriptive statistics to identify missing data, data errors, outliers and check for data imbalance in the outcome variable
2. Resolve any data errors or issues
3. Exploratory data analysis:
    a. Create crosstabs or plots to explore the relationship between each variable and the outcome (retention) to understand patterns in the data
    b. Create crosstabs or plots to explore the relationship between each variable and the other variables. Assess statistical significance of any correlated variables and remove any redundant variables.
4. Split data into training (80%) and test (20%) datasets
5. Train a decision tree classifier using the training dataset
6. Test the decision tree classifier on the test dataset to evaluate model performance using accuracy, precision, recall, and F1 score
7. Plot the decision tree model and assess information gain at each split in the tree to identify the variables that have highest importance in classifying the outcome

Programming language(s):

Data task	Programming language(s)	Packages
Extracting data	SQL	n/a
Exploratory Data Analysis	R or Python	ggplot or matplotlib + seaborn
Machine Learning	R or Python	caret or scikit-learn

Measuring Success:

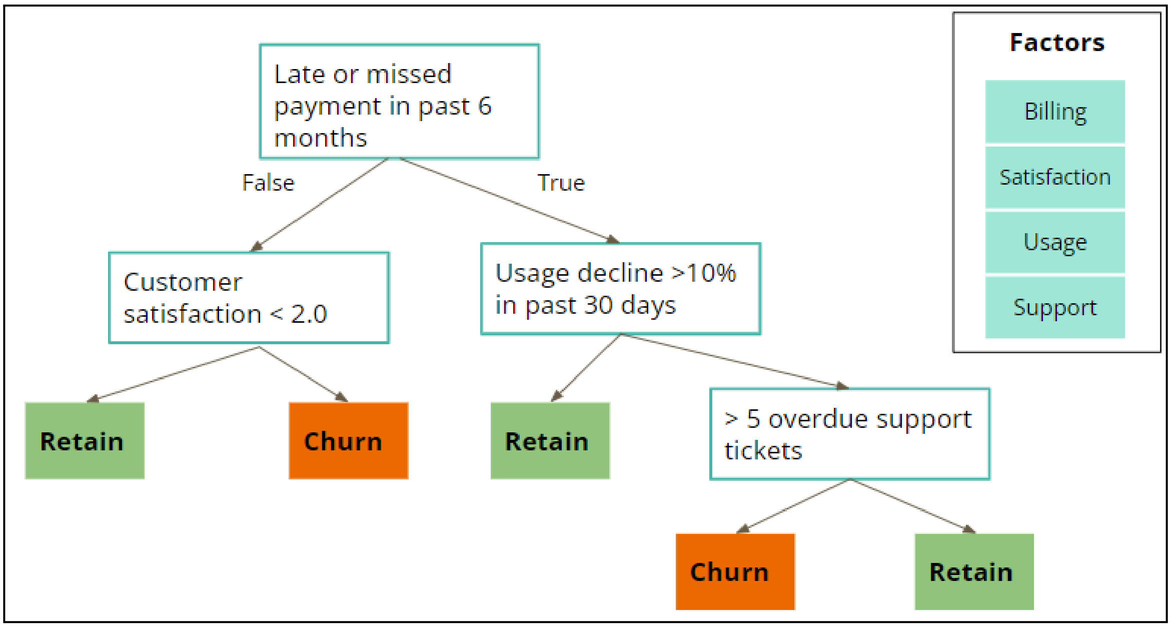
Success would be measured in the following ways:

- The model should be able to accurately predict churn in the test dataset as measured by overall accuracy and F1 score
- The model identifies the most important predictors of customer churn and retention as measured by information gain
- The identified predictors and decision rules inform the creation of actionable deliverables such as customer health models and reports
- The process for data gathering, exploration and modeling is clearly documented and stored in a centralized repository
- Any data issues are clearly documented and reported for investigation and resolution by the appropriate team

Part 3:

Task: Mock-up reports that convey actionable insights from retention exploration to executive leadership

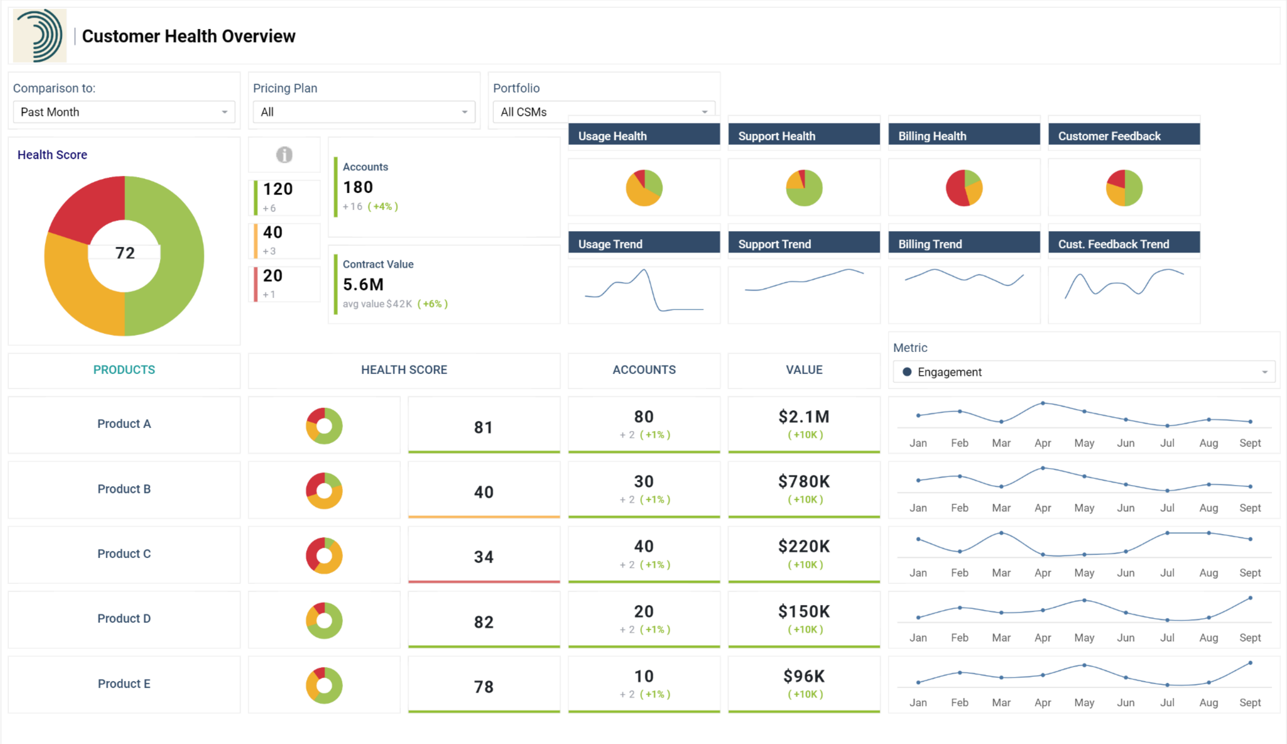
The decision tree model, the outcome of Part 2, would be visualized (see example below) to provide stakeholders with an understanding of the most important factors and specific rules that predict retention and churn at the company:


Late or missed payment in past 6 months
False
True
Customer satisfaction < 2.0
Usage decline >10% in past 30 days
Retain
Churn
Retain
> 5 overdue support tickets
Churn
Retain
Factors
Billing
Satisfaction
Usage
Support

These decision rules (e.g. “usage decline > 10% in the past 30 days”) would inform the creation of a series of customer health scores. For example, a “Usage health score” would incorporate all the elements from the decision tree that relate to a customer’s use and engagement with the company’s products. These health scores would be combined to provide an overall health score for each customer and then rolled-up by portfolio, product, segment, or the entire customer base.The decision rules would also facilitate the setting of targets or benchmarks against which current data can be evaluated; providing a useful visual comparison in dashboards.

The example dashboards (below) focus on the visualization of customer health scores to give leadership actionable information to maximize customer retention by allowing them to quickly identify trends in overall health by segment or product, compare the health of different factors and see individual customer data in relation to established targets and benchmarks. These dashboards are actionable because they provide an early warning system to target specific segments, customers, or factors to mitigate churn *before* churn occurs. It would also be important to provide leadership with a way to track subsequent intervention efforts and retention outcomes.

Example Customer Health dashboards [Customer Health Dashboard.pdf]
Breakdown by Product:



Customer Health Overview

Comparison to: Past Month

Pricing Plan: All

Portfolio: All CSMs

Health Score: 72

120	+6
40	+3
20	+1

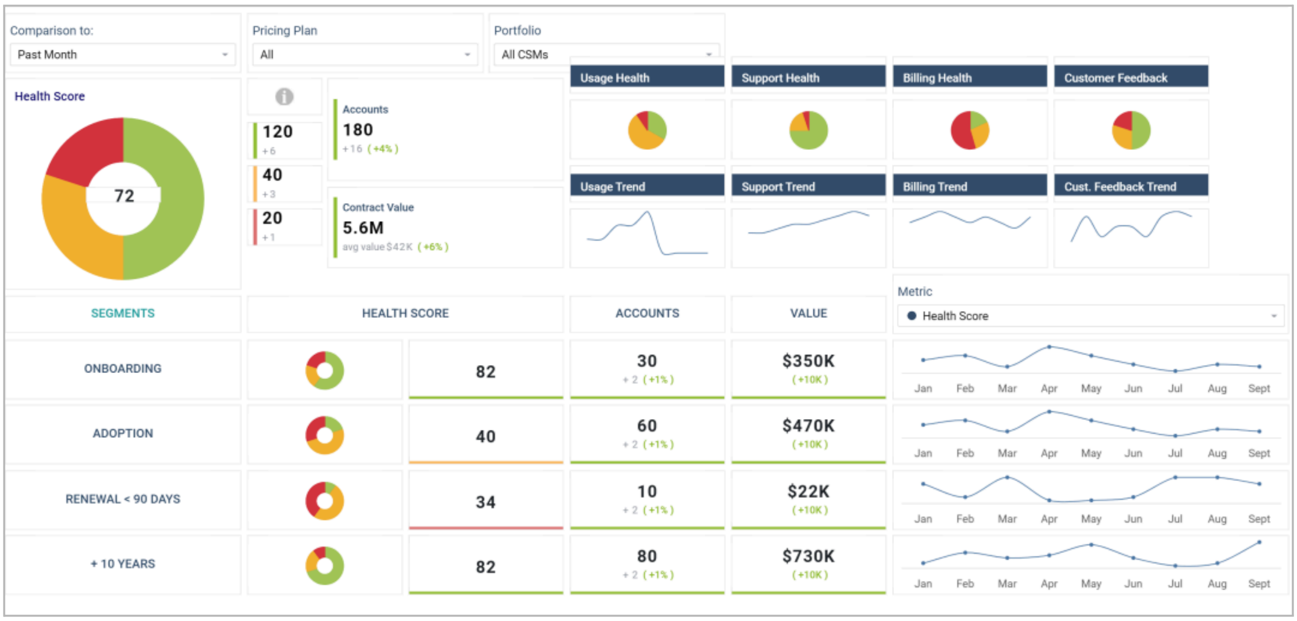
Accounts
180
+16 (+4%)

Contract Value
5.6M
avg value $42K (+6%)

Metric: ● Engagement

PRODUCTS	HEALTH SCORE	ACCOUNTS	VALUE
Product A	81	80 + 2 (+1%)	$2.1M (+10K)
Product B	40	30 + 2 (+1%)	$780K (+10K)
Product C	34	40 + 2 (+1%)	$220K (+10K)
Product D	82	20 + 2 (+1%)	$150K (+10K)
Product E	78	10 + 2 (+1%)	$96K (+10K)

Breakdown by Segment:



Comparison to: Past Month

Pricing Plan: All

Portfolio: All CSMs

Health Score: 72

120	+6
40	+3
20	+1

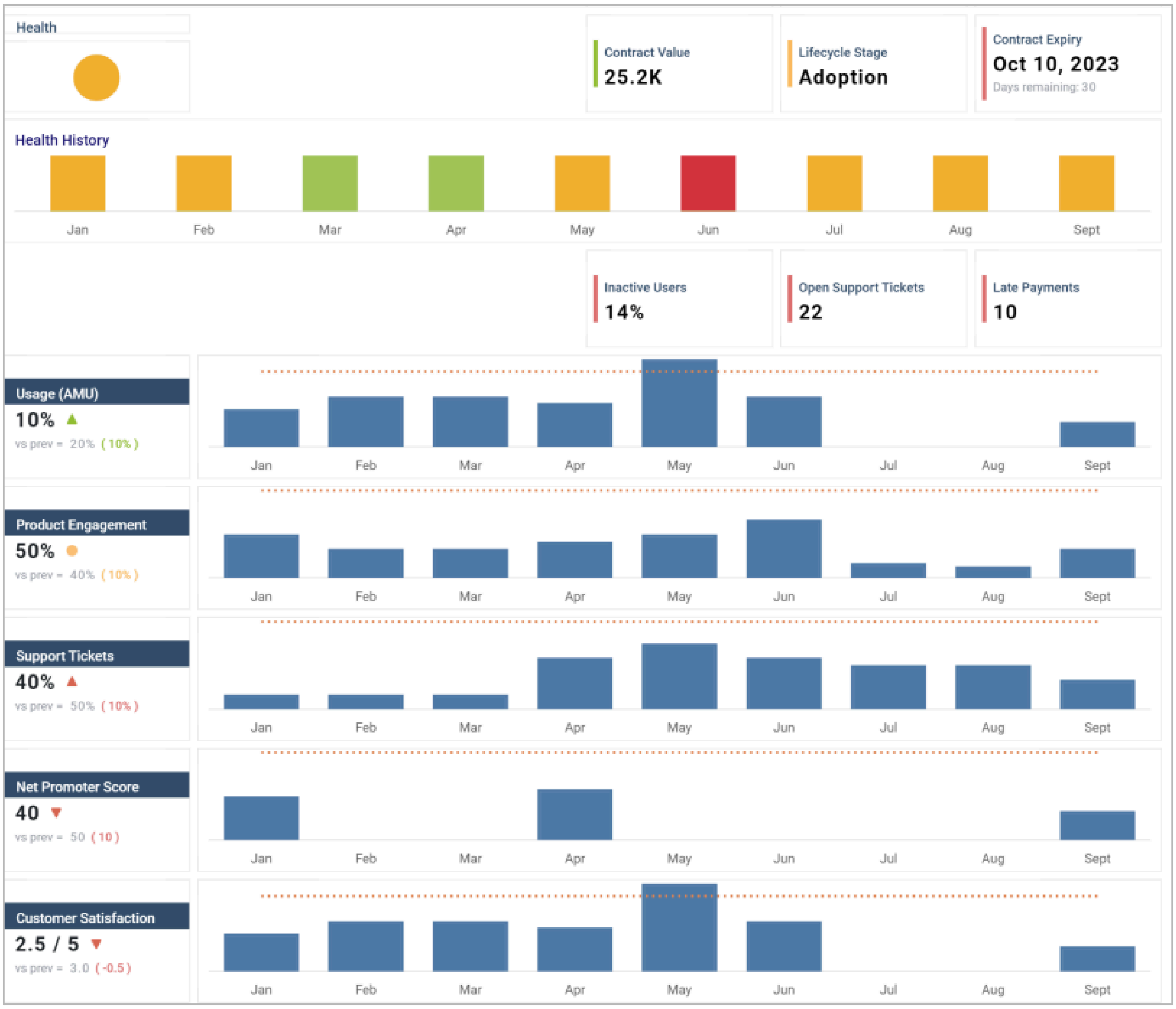
Accounts
180
+16 (+4%)

Contract Value
5.6M
avg value $42K (+6%)

Metric: ● Health Score

SEGMENTS	HEALTH SCORE	ACCOUNTS	VALUE
ONBOARDING	82	30 + 2 (+1%)	$350K (+10K)
ADOPTION	40	60 + 2 (+1%)	$470K (+10K)
RENEWAL < 90 DAYS	34	10 + 2 (+1%)	$22K (+10K)
+ 10 YEARS	82	80 + 2 (+1%)	$730K (+10K)

Individual Customer Health Dashboard:



Health

●

Contract Value
25.2K

Lifecycle Stage
Adoption

Contract Expiry
Oct 10, 2023
Days remaining: 30

Health History

Inactive Users
14%

Open Support Tickets
22

Late Payments
10

Usage (AMU)
10% ▲
vs prev = 20% (10%)

Product Engagement
50% ●
vs prev = 40% (10%)

Support Tickets
40% ▲
vs prev = 50% (10%)

Net Promoter Score
40 ▼
vs prev = 50 (10)

Customer Satisfaction
2.5 / 5 ▼
vs prev = 3.0 (-0.5)