Technical Report

an analysis of Quark VPN customer churn

In this report, you will find details of the methodology, findings, and conclusion from an analysis of 2 years data from VPN service Quark. The main goal was to identify what's driving customer churn and provide actions to improve retention.

▼ Methodology and data preparation

1. Load and clean the data

There are three datasets provided (`vpn_customer_data.csv`, `vpn_events.csv`, `vpn_external_factors.csv`). A quick inspection shows that some cleaning and transform tasks were necessary:

- All date related columns were converted from string object to date time object.

2. Feature engineering

A few key features were created to represent user behaviour for analysis:

- `is_churned`: This is the target variable.
- `tenure_days`: The duration of the user's subscription lifetime in days.

  Churn users: churn date - sign up date

  Active user: last churn date - sign up date
- `cost_per_hour`: The `plan_price` for annual users was normalised to a monthly basis (price /12). Normalised price was then divided by `monthly_usage_hours`.
- Parsing categorical columns: `num_ticket_types`, `num_devices`, `has_technical_ticket`, etc

3. Merge data

Customer data was left joined with the external factor data on `country` and `year_month`. Then event data was joined for a later analysis to look into acquisition source.

▼ Exploratory analysis and some findings

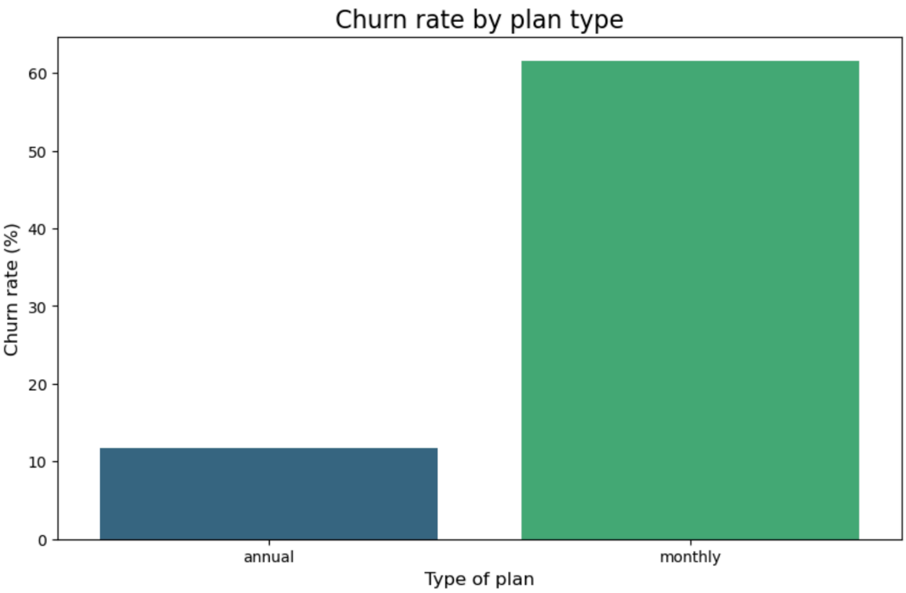
I started with a broad exploration for the most primary churn factor and then dig deeper into when and why the user churned.

1. The main factor: Monthly vs Annual subscription

- Thought process: Monthly users have multiple renewal points where they can consider canceling the subscription. The less user commitment tends to lead to higher churn rate.
- Findings: This hypothesis was confirmed. The analysis showed a huge difference in churn rate.

> 💡 Monthly subscribers churn rate is 61.6% which is 5 times higher than that of annual subscribers (11.7%)

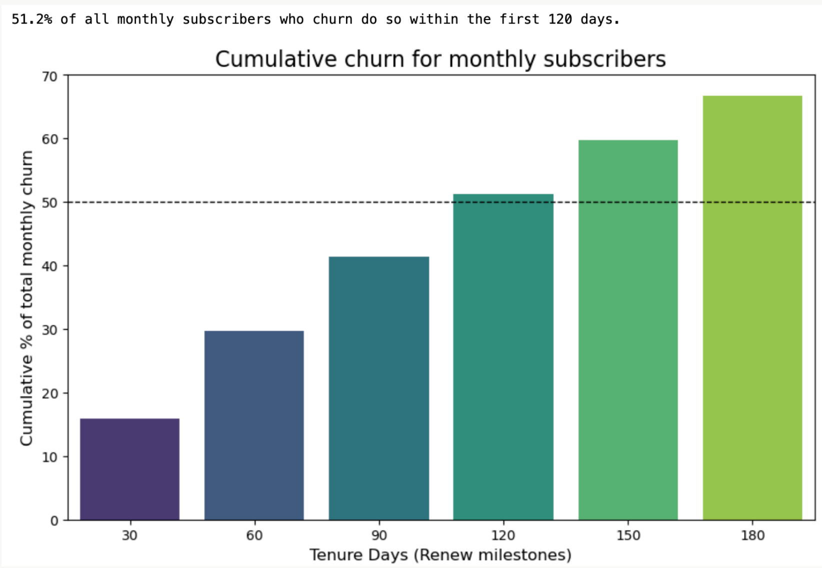
The business's churn problem is mostly happening within the monthly user segment, and therefore was the primary focus of the rest of the analysis.



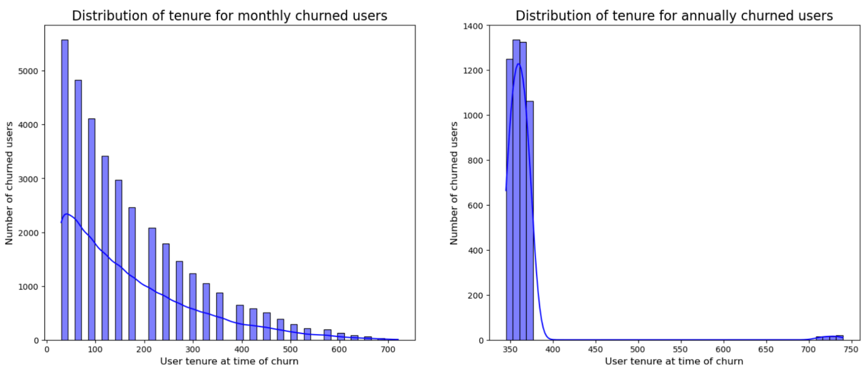
2. "When" did the users churn

- Thought process: after identifying the "who", the next step was to see "when" they leave. The hypothesis was that churn is a "slow death" situation where we see users churning away each month instead of a "instant drop" in the first month.
- Findings:
    - Monthly churn in early stages for monthly subscribers

      Users churned at renewal points, and 30 days after subscription topped the list where we see more than 15% of the monthly churned subscribers left.

> 💡 A cumulative analysis also showed that over 50% of all all monthly users who ever churn are lost within the first 4 months.
>
> 
>

- Annual renewal churn occurs right around the Day 365 mark.



3. "Why" did the user churn

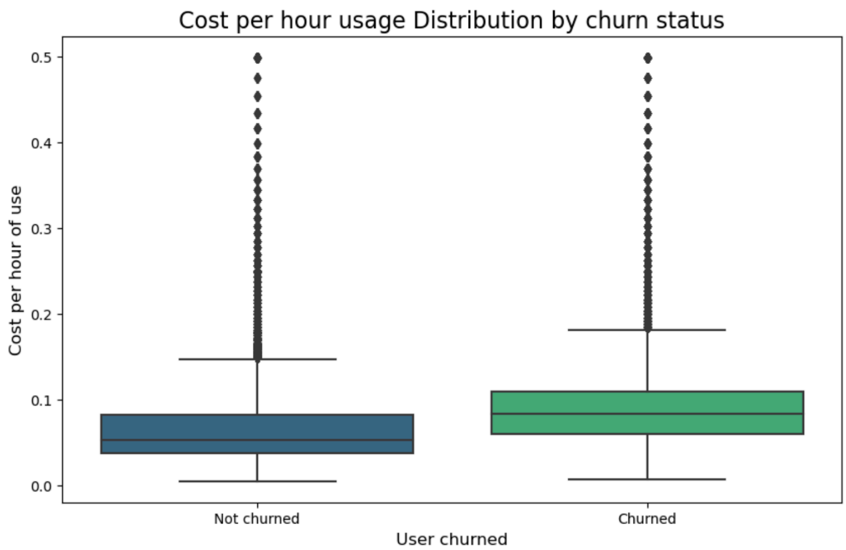
- Thought process: There were two parts established, the user segment and the timing of churn. Here the focus was to identify the reasons for cancellation.
- Findings:
    - Bad value for money is a good indicator

      Hypothesis: Users has the perception that the VPN service does not worth the recurring cost, and this can be quantified by the cost per hour or usage.

      Analysis: The metric created `cost_per_hour` is a good predictor. The box plot here showed the median cost for churned users was 58% higher than the non-churned users. A t-test also confirmed the difference in mean was highly statistically significant, showing that a poor value perception is a great indicator for churn.

```
              count      mean       std       min       25%       50%  \
is_churned
0           59957.0  0.064767  0.044558  0.005010  0.037803  0.053085
1           40043.0  0.091234  0.053232  0.007253  0.060545  0.083950

                 75%     max
is_churned
0           0.081803  0.4995
1           0.109780  0.4995
```



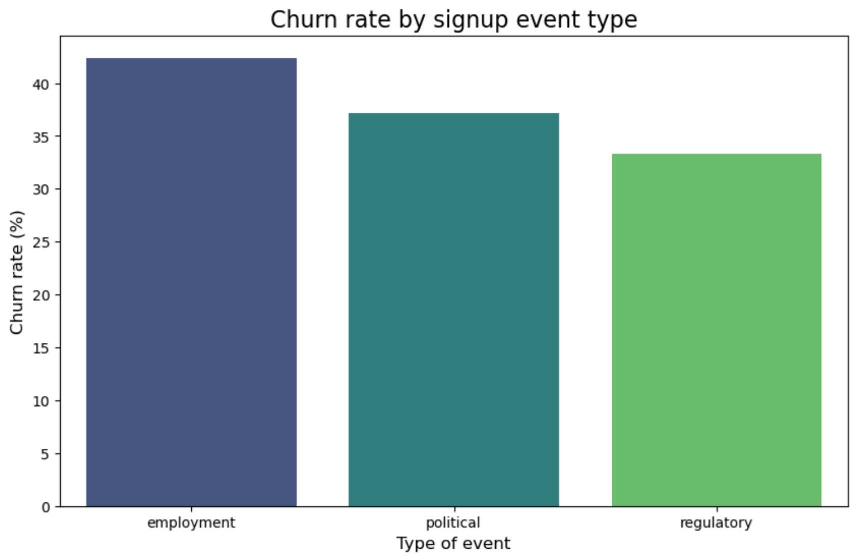
    - Acquisition source of event-driven users is a good predictor for user quality

      Hypothesis: Some real world events that drove users to sign up will impact their retention due to the initial different motivation.

      Analysis: The type of the events was a good filter. Users who signed up during months with "political" (33%) or "regulatory" (37%) events had a lower churn rate comparing that of "employment" (42%)

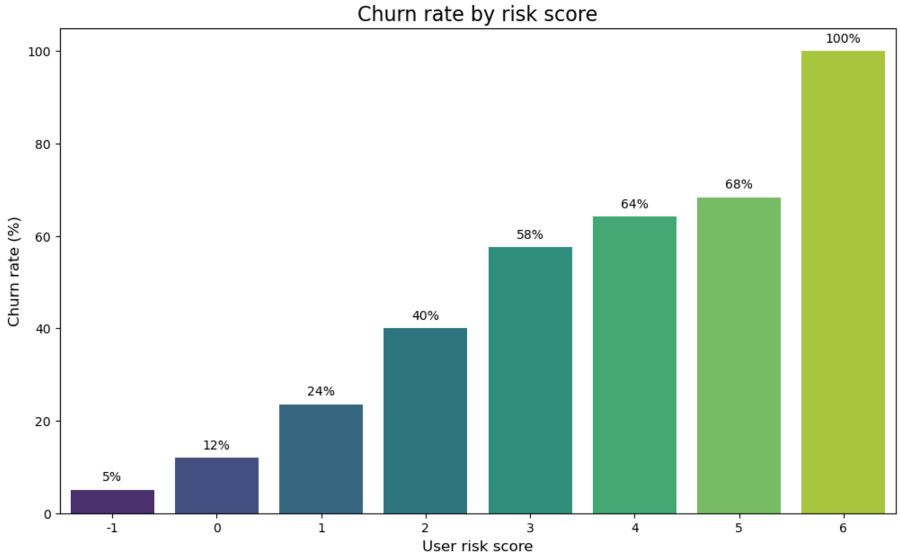
events.

```
Churn rate by event type
   event_type  is_churned  churn_rate_percentage
0  employment    0.423967              42.396732
1   political    0.371289              37.128874
2  regulatory    0.332948              33.294798
```



4. Build an at-risk profile

- Thought process: The exploratory analysis identified several drivers for the churn. The last step was to combine these separate signals into a score to identify at risk users. I thought about using machine learning models like logistic regression to predict if the user is going to churn or not. But I feel that a score model would be better to interpret to a non-technical audience. This will also give us a guide to "who" we should target first to improve user retention.
- Each user was assigned points based on the following rules:
  - User commitment (max 2 points):
    - `+2 points` if on a monthly plan with tenure ≤ 30 days (Highest Risk).
    - `+1 point` if on a monthly plan with tenure > 30 days.
  - Value for money (max 1 point):
    - `+1 point` if their `cost_per_hour` was above the population median.
  - Engagement (max 2 points):
    - `+1 point` if they connected to fewer countries than the median.
    - `+1 point` if their `monthly_usage_hours` was below the median.
  - Acquisition source (max +1 or -1 point):
    - `+1 point` if they signed up during an `employment` event.
    - `1 point` if they signed up during a `political` **or** `regulatory` event.
- Findings: As the risk score increases, the probability of churn rises dramatically, from just 5% to 100% for the highest risk users. This showed that while some individual signals can be weak, when combined, they provided a clear and also actionable way to segment the user base.



▼ Hypotheses ruled out

During the analysis, several other logical hypotheses were tested and found not to be significant drivers of churn in the dataset. These included:

- The number and types of support tickets raised
- The number and types of devices used
- The user's country and other external factors

▼ Final Conclusion

From the above analysis, we found that there are two user bases: the loyal annual subscribers and the high-risk monthly subscriber. Churn for the latter group is a persistent problem, and that is identifiable by a bad perception of value and concentrated at the first few months of renewal milestones. With the risk score model, we are able to take actions based on their scores:

- Low risk users (score 0-1 / churn rate ≤ 25%): we should reward and nurture these users. There is no need for discounts and maybe focus on early access to new features, or simple thank you email to reinforce them choosing us.
- Medium risk users (score 2-3 / churn rate ≤ 60%): these users are on the fence of leaving the subscription plan. we should engage and educate them to improve value.
    - For instance, we launch email campaigns highlighting different use cases of features they have access to (i.e. Tools we provide you to combat these service providers OR What you can do to avoid surveillance when living in these countries).
    - What we can also do is to offer a one-time discount for them to upgrade to an annual plan, that way they can "get more value and save money".
- High risk users (score 4+ / churn rate >60%): Aggressive retention offers is probably better than standard engagement. This segment needs proactive and high incentive offers to prevent them cancelling the plan.
    - For instance, a "last chance" email offering them a discount (50% off your next three months) to break the churn cycle

By implementing this tiered and targeted strategy, retention resources can be allocated more effectively.