Task 1
Business Understanding & Hypothesis Framing
Understanding the business context and problem statement
Author Prateek Naharia 06/15/2023

Subject: Customer Churn Prediction & Mitigation Strategy at PowerCo

Hi Associate Director,

I hope this email finds you well.

As per the discussion on PowerCo's SME customer churn, here is a detailed approach to test the hypothesis that price sensitivity is driving SME customer churn.

Our Hypothesis: There is a significant correlation between price changes and the rate of customer churn among PowerCo's SME customers.

I would like to take this further in 3 Phases - the following hypotheses :

Phase 1. Data Collection & Key Factors : The data can be sourced from the PowerCo's internal systems like CRM, billing systems, customer feedbacks from the official channels and social media platforms, external sources like market research or public datasets (for eg. on geopolitical issues, geography, electricity/gas prices).

Considering PowerCo's internal systems: This will be based on the historical data on churn, service usage, customer behaviors, price variation and subscription data .
Moreover this would also include, customer feedback data, price variations during the customer's subscription period, customer demographics (information about SME, type & location), billing data (monthly, quarterly & yearly) consisting of other cost factors like electricity & gas usage.
There can be other fields that could be used to explore like competitor pricing data, customer tenure, feedback/satisfaction data.
Each of the above fields could help shed light on various factors that could contribute to a customer's decision to churn.

Dataframe : Ideally the above mentioned factors should be represented as columns and rows as customer/ individual SME { (individual SME(rows) x factors(columns)) }
Here's a snapshot of potential columns:
Cust_Id, industry, tenure, location, size, avg_monthly_usage, total_billing_amout, num_price_change, last_price_change, num_queries_complaints, num_resolved_complaints, feedback_score, churn_status, churn_reason.

The above can be a mix of binary, text, numerical columns & we can increase the number of columns by feature engineering - eg complaint resolution rate etc.

Phase 2 : EDA
Exploratory Data Analysis can be conducted to identify anomalies, uncover patterns, trends and data relations.

- Statistical analysis for each column, including count, mean, mode, median, min, max & standard deviation - data distribution sense.
- Correlation Analysis between different variables & visualizations (heatmap of correlation matrix).
- Null/Missing value analysis & null value imputational scope.
- Check outliers which might skew the results, using boxplots or scatterplots.
- Churn Rate Analysis (CRA), churn rates x price variation trends, patterns & relationships.
- Customer segments price variation sensitivity analysis.
- Time Series Analysis - customer churn seasonality/seasonal trends.
- Customer Interaction Analysis
- Analyze customer churn rates across different segments of customers.
- Sentimental Analysis or text data analysis for understanding the main reason for churn.
- Feature Engineering across the fields for better understanding.

Phase 3 : Model Building & Evaluation
This is a binary classification problem where the outcome (Churn) would be - { Yes | No n\}
The following procedure can be followed :
Data Preprocessing -> Feature Selection -> Split DataSet -> Model Selection & Training -> Model Evaluation -> Hyperparameter Tuning -> Validation & Interpretation -> Deployment -> Monitoring (Monitor the model over the time as with new data the performance degrades).

Analytical models that could be considered for the problem
- Logistic Regression - Binary classification problem (churn Yes or No Prediction).
- Neural Networks: This consists of multiple layers of neurons that transform the input features into higher-level representations that are useful for predicting churn.
- SVM (Support Vector Machine): The model can try to find the hyperplane in the feature space that best separates the churners from the non-churners, based on the closest points (the support vectors).

To implement the discount strategy, the model output can be used to identify the customers who are most at risk of churn and may therefore benefit from a 20% discount offer.