Customer Churn Prediction

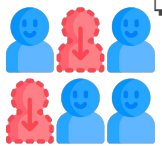
Business Scenario

- With our new TV app Southeast Asia Launch right around the corner, we want to ensure the rollout is successful and **profit making** for parent company
- **Subscription plan**: Our app offers a subscription plan where customers can sign up for a 1-month or 3-month plan. After 30/90 days, customers will be automatically charged based on the package selected. That is, the **initial 30/90 days trial** period is **free**

Photo by CHUTTERSNAP on Unsplash

How can we ensure that our customers, stay longer with us, beyond their free trial?

Success Metrics



- Given our subscription based business model, we could start with a target metric like:

  Customer churn rate

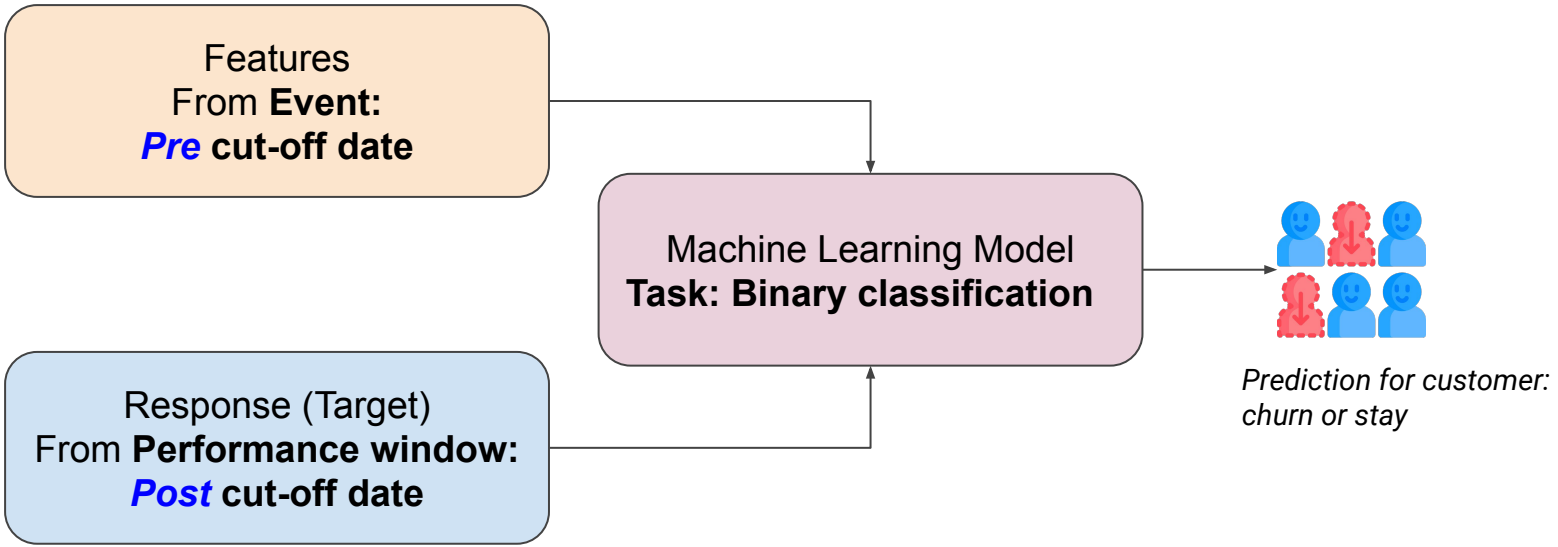
- Understanding if a customer might churn based on what we can glean about them during their lifetime value, could give us a **leading indication to retain** them before a churn could actually happen

- With such early indication, we could target campaign management measures to proactively engage with highest-risk-of-churning customers
  - *Example, LinkedIn sends discounts to continue being a **paid** Premium member when free trials are close to ending*

***Measure** of success: **Reduce** customer **churn** → **Increase** lifetime value → **Increase** revenue (cost of retaining existing customer is lesser than acquiring new!)*

Applying Data Science: Predicting Customer Churn

- Our predictive model should be able to answer these questions:
  - "Is a given customer going to churn within the **next X months**?"
  - "How early do we want to **intervene with actions before a churn** can actually happen?"



Modeling Framework: Predicting Customer Churn

Customer Churn Model: Business Application

Monthly **Active Customers**

Machine Learning Model Trained with **Customer Churn History**:

Features from **Event** period
+
Response from **Performance window** period

Churn Score **LOW**

Churn Score **HIGH**

No actions required

Activate **Campaign Management**

Data Understanding

Background

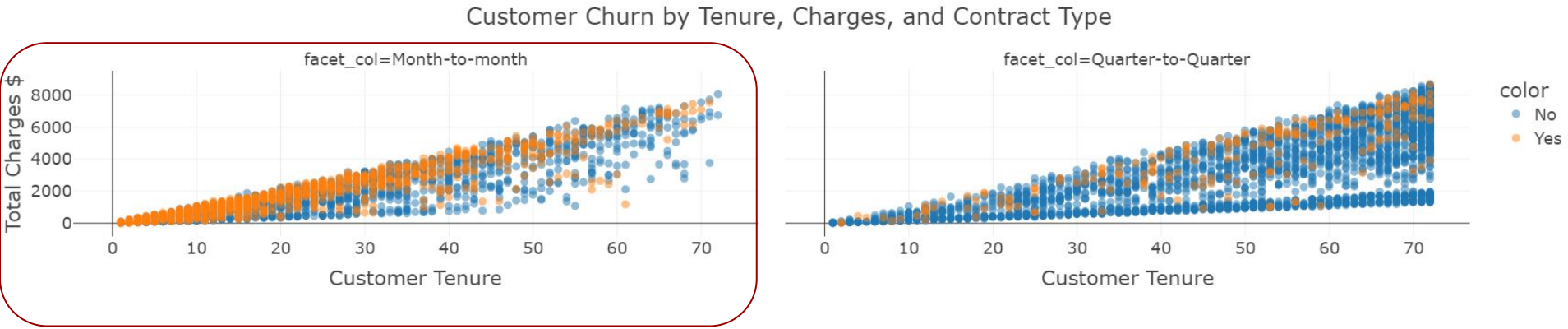
- For the purposes of technique illustration, an open-source Telecom Customer Churn dataset has been used
- Though some of the fields can be irrelevant to our use case, we will assume this is proxy data from our app's Global Customer base from countries where the service already exists
- We can continue refining and localising our model once we have region-specific customer base post launch

Transformations

Variables	Original	Transformed	Description
TotalCharges	Some noisy string values	- String replacement with nulls - Nulls imputed during modeling	Captures total charges to customer
Contract	“Month-to-month”, “One year”, “Two year”	“Month-to-month”, “Quarter-to-Quarter”	Captures contract term of customer - Tailored to fit subscription plan per our business model to capture monthly & quarterly subscriptions
StreamingMovies	“Yes”, “No”, “No internet service”	“Yes”, “No”	Let’s assume that these represent if a customer streams Movies on our platform - “No service” isn’t relevant (it is present originally as this is a Telco dataset)
PhoneService, MultipleLines, InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV	Present	Dropped	These variables are more specific to Telco service, given the nature of this dataset and are thus dropped for our business context

Exploratory Data Analysis

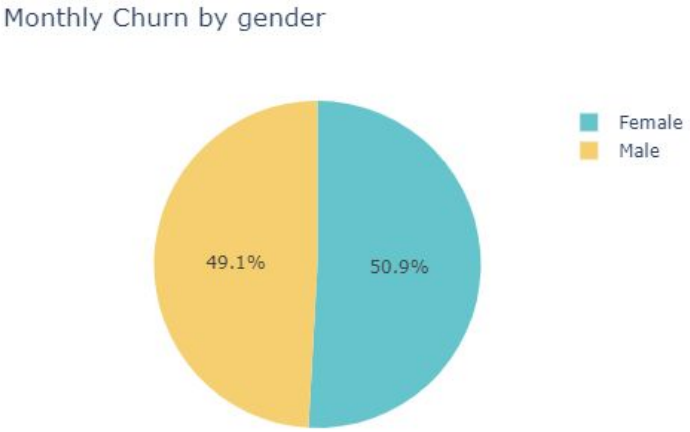
Churn by Tenure, Charge, Contract

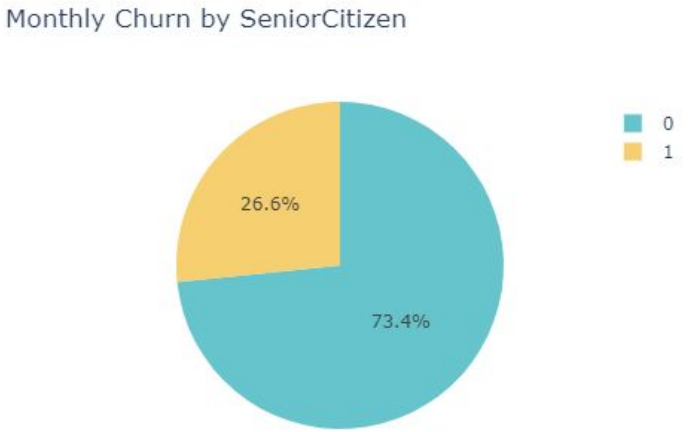


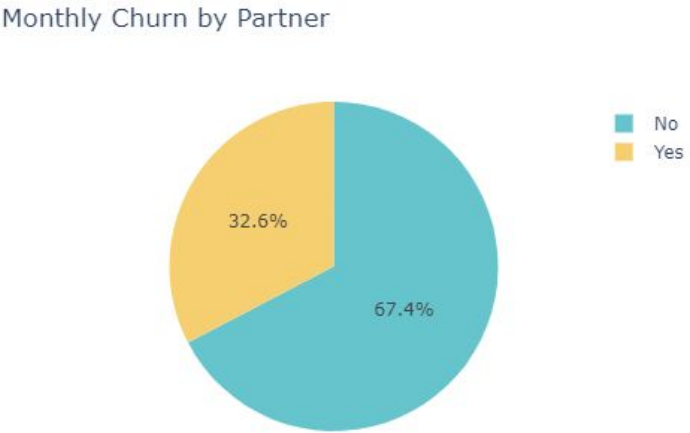
Above visualization captures Customer Churn by the 2 types of Contract: 'Month-to-month' and 'Quarter-to-Quarter', in relation to Pricing and Tenure:

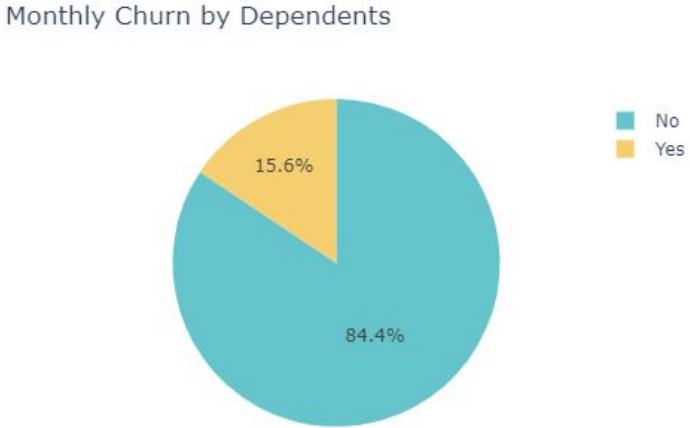
- ❏ 'Month-to-month' churn is significantly higher *(more orange dots)* as we would expect. This group of customers refrain from long-term commitments, thus necessitating more focus for retention measure application

Churn by Customer *Profile* Attributes





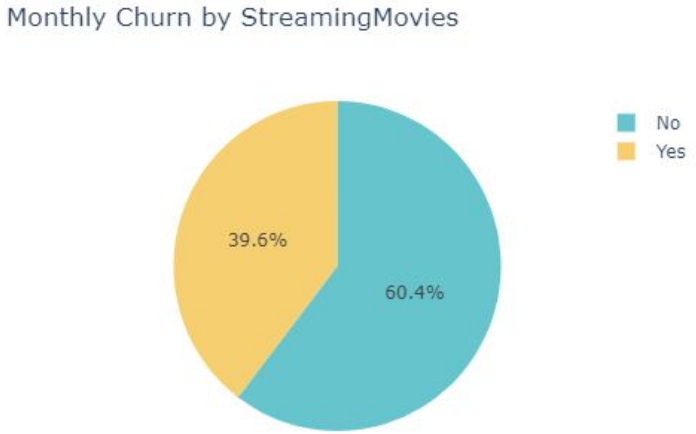


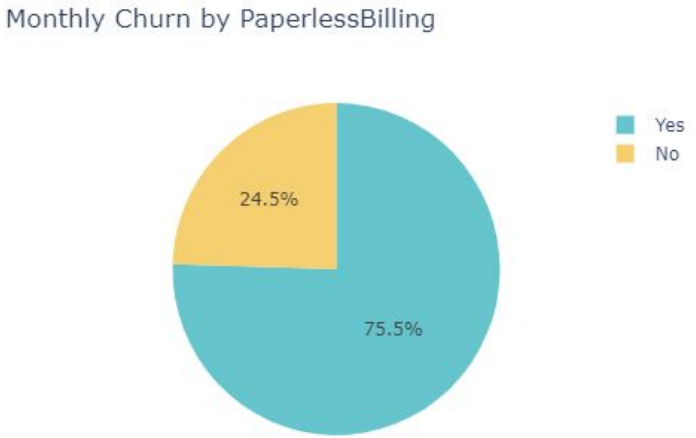


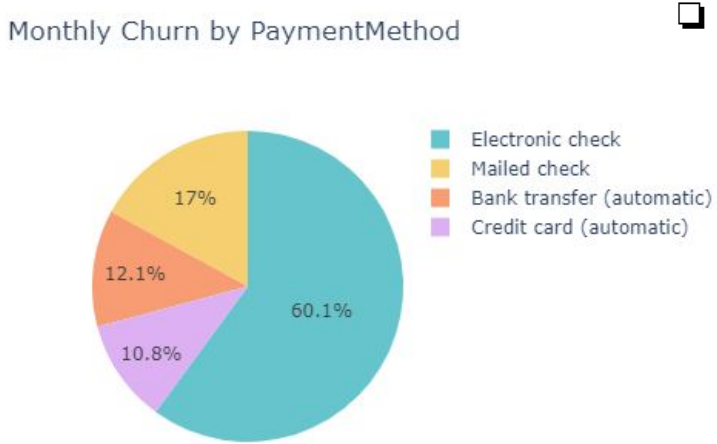
Within the monthly subscription group that churned, here are insights by various customer profiles:

- ❏ Churn within genders was comparable
- ❏ Higher% customers that churned were younger & single population

Churn by Customer *Subscription* Attributes



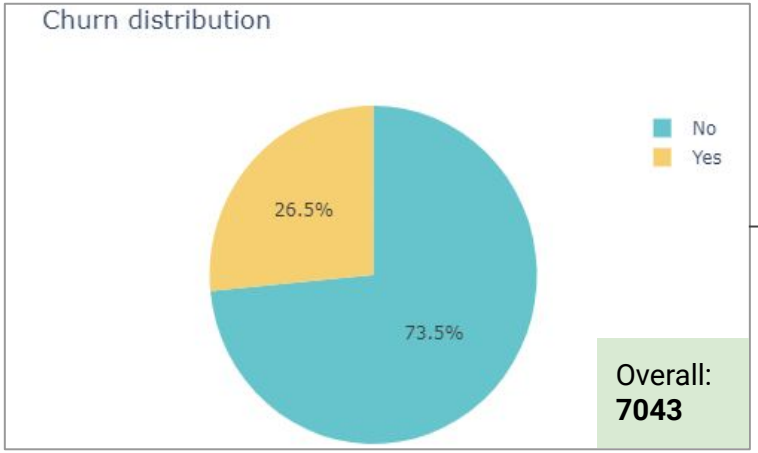




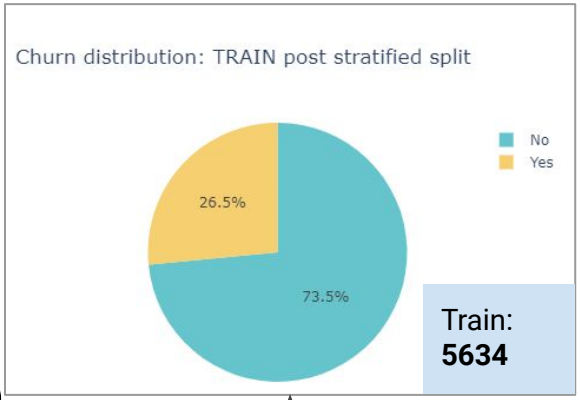
Within the monthly subscription group that churned, here are insights by various customer subscriptions:

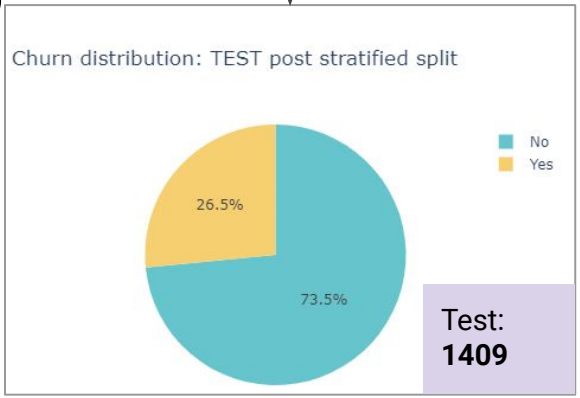
- Customers that watched movies on our app churned more vs those that didn't
- Higher% customers that churned had opted for Paperless Billing and this coincided with higher churn from online payment mode of Electronic check

Train-Test Split



Train-test stratified split : 80-20%

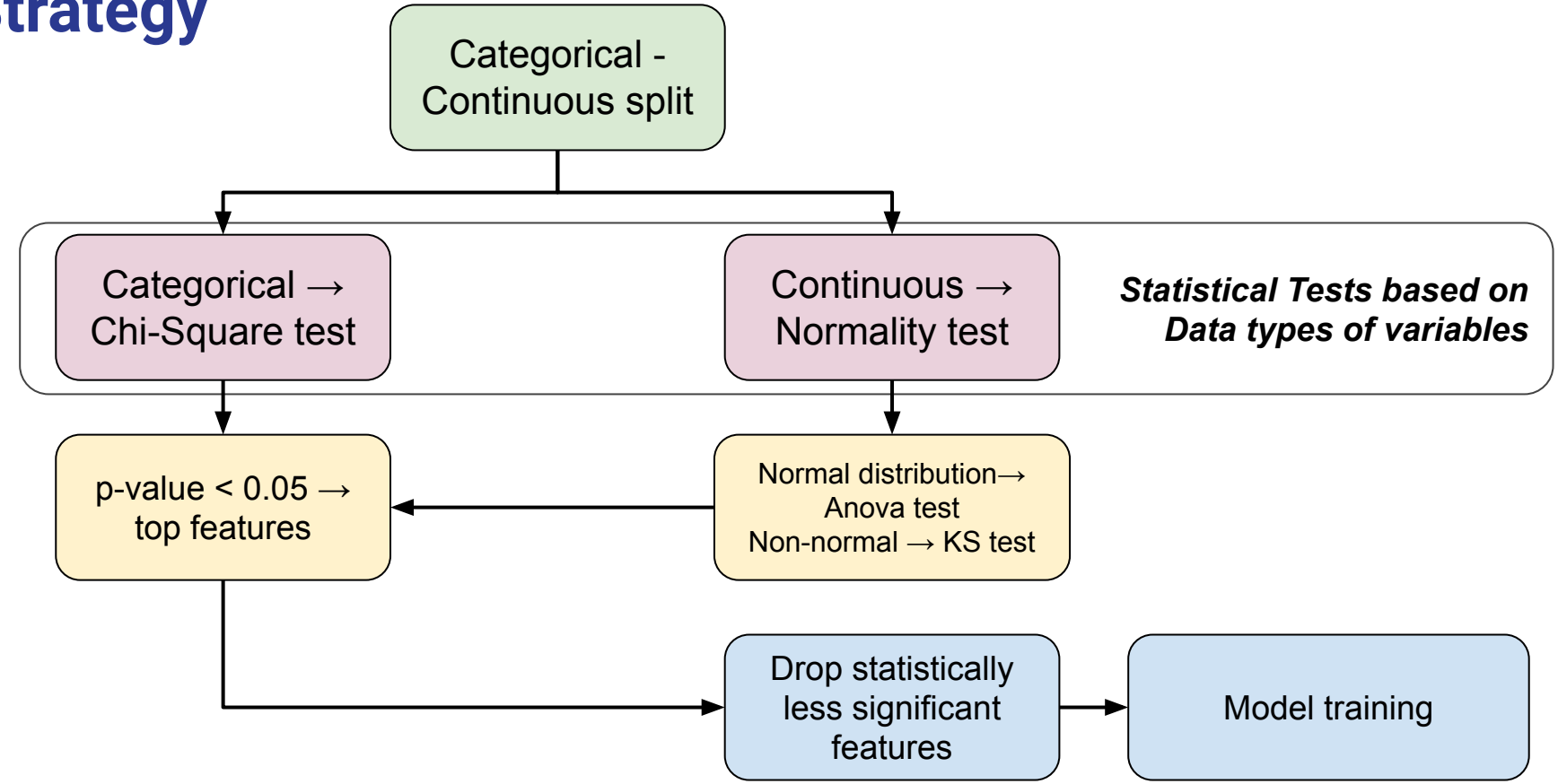




- As we would expect, Churn data is imbalanced in distribution
- It is crucial to maintain the sample% between the churned and not-churned in the same proportion when we split data for model training and testing
- This is achieved by doing a stratified split as as shown above, maintains the same proportion to closely mimic parent dataset

Features Selection

Strategy


Categorical - Continuous split
Categorical → Chi-Square test
Continuous → Normality test
Statistical Tests based on Data types of variables
p-value < 0.05 → top features
Normal distribution→ Anova test
Non-normal → KS test
Drop statistically less significant features
Model training

Results

Categorical features with most statistically significant impact on “Churn”

	p_value
Contract	6.911966e-114
Dependents	2.450103e-24
PaperlessBilling	5.513289e-21
Partner	1.116854e-15
PaymentMethod	4.731862e-11
StreamingMovies	4.363247e-04
gender	4.581647e-01

< 0.05

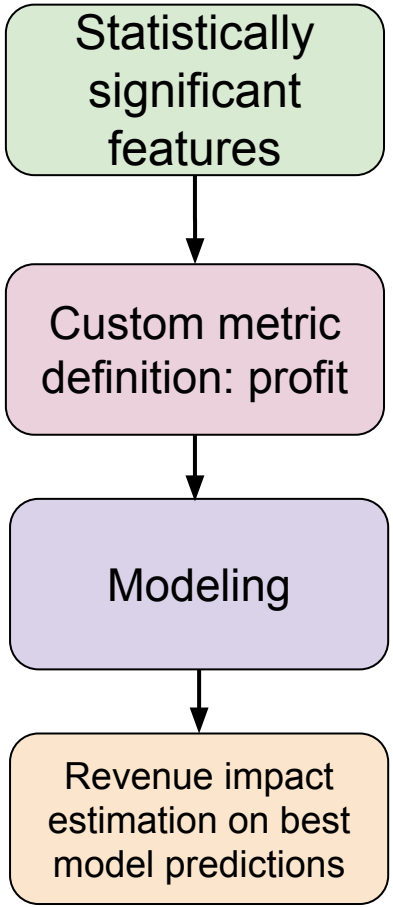
Continuous features with most statistically significant impact on “Churn”

	p_value
tenure	9.065749e-123
MonthlyCharges	3.527176e-61
TotalCharges	2.948654e-48
SeniorCitizen	5.880770e-16

- Top **categorical** features highly affecting customer Churn based on p-value <0.05
  - All except Gender as previously confirmed in EDA
    - This means we could potentially drop Gender as a feature during model training as it does not have any statistical impact on Churn
- Top **continuous** features highly affecting customer Churn based on p-value <0.05
  - All 4: Tenure, Pricing variables, SeniorCitizen

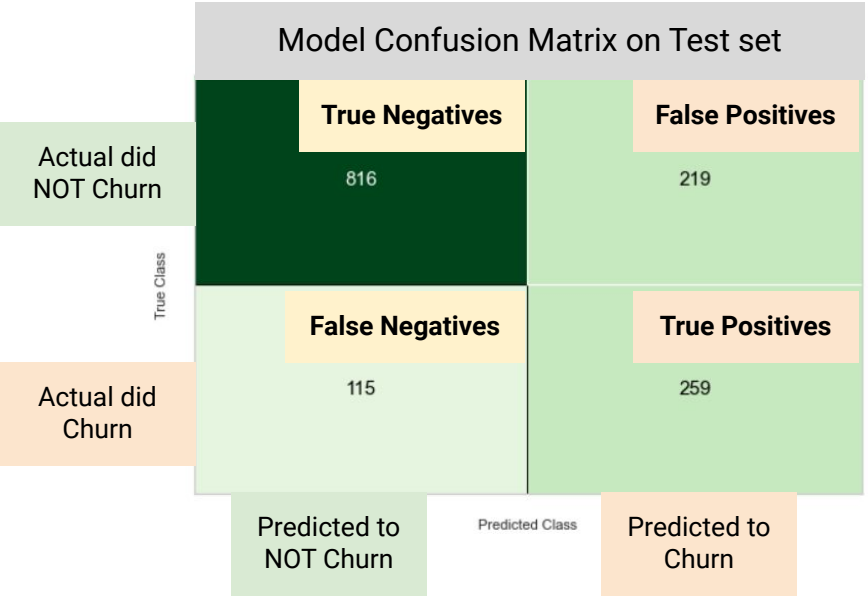
Modeling

Strategy



- Based on our business context, we ultimately want to prevent churn so we can contribute to company's revenue growth
- We can incorporate this to access our churn model, because, the reward of ***true positives*** is way different than the cost of ***false positives***
- We can thus custom-define a profit metric to evaluate the revenue impact/savings from our model's churn prediction
  - **Assumptions:**
    - $1,00 voucher will be offered to all the customers identified as churn (true + false positives)
    - Stopping a churn translates to $1,000 gain in Customer Lifetime Value (CLV)

Results



Model churn predictions			$		Total
Predicted churn that actually churned	True positives	**259**	Gain in CLV for stopping a churn	**1,000**	259000
Total predicted churn	True + False positives	**478**	Voucher discount for identified churn	**100**	-47800
Savings:					**$ 211,200**

- The matrix *(called confusion matrix)* beside is an output from the machine learning prediction model on the **1,409** test data points
- The 259 **True Positives** (18%) are the customers *predicted by the model to churn that actually did churn* → without the model, we would not have been able to catch this churn. The predictions just helped us extend lifetime value for these customers!
- The 219 **False Positives** (16%) are the customers *predicted by the model to churn that actually did NOT churn* → this is where we'll lose money as we would have offered them discount voucher, but they were anyway going to stay
- The profit calculation based on the above 2 + weaving in $ values spent by company on voucher, CLV, results in **~200k worth $$ savings!**
- 816 **True Negatives** (60%) are customers *predicted to stay and they did actually stay*; while 115 **False Negatives** (8%) is missed opportunity → the model missed to catch this 8% that actually churned

Inferences

Recommendations & Future Work

Recommendations

- Recommendation engine to recommend content on platform catered for younger-single population (explore adding ‘Age’ for new customer sign-up)
- Review movies that are currently streaming and update to new releases, since this has proven to drive revenue in the US
- Engage with customers paying through Electronic check payment mode & encourage transfer to other online transfer modes
- Marketing campaigns targeted at the same population, since they historically have higher chance of churn

Future Work

- Improve False Negatives on model
- NLP sentiment, topic prediction on churning customer text feedback from exit survey
- In the absence of sensitive fields like Gender during customer sign-up, we can do name-to-gender predictions to still populate the gender feature for modeling

References

- Icons: https://www.flaticon.com/
- Dataset: https://www.kaggle.com/blastchar/telco-customer-churn
- Link to code notebook: https://nbviewer.org/github/shilpaleo/customer_churn_prediction/blob/main/notebooks/customer_churn.ipynb?flush_cache=true

Thank you!

Photo by Yulia Matvienko on Unsplash