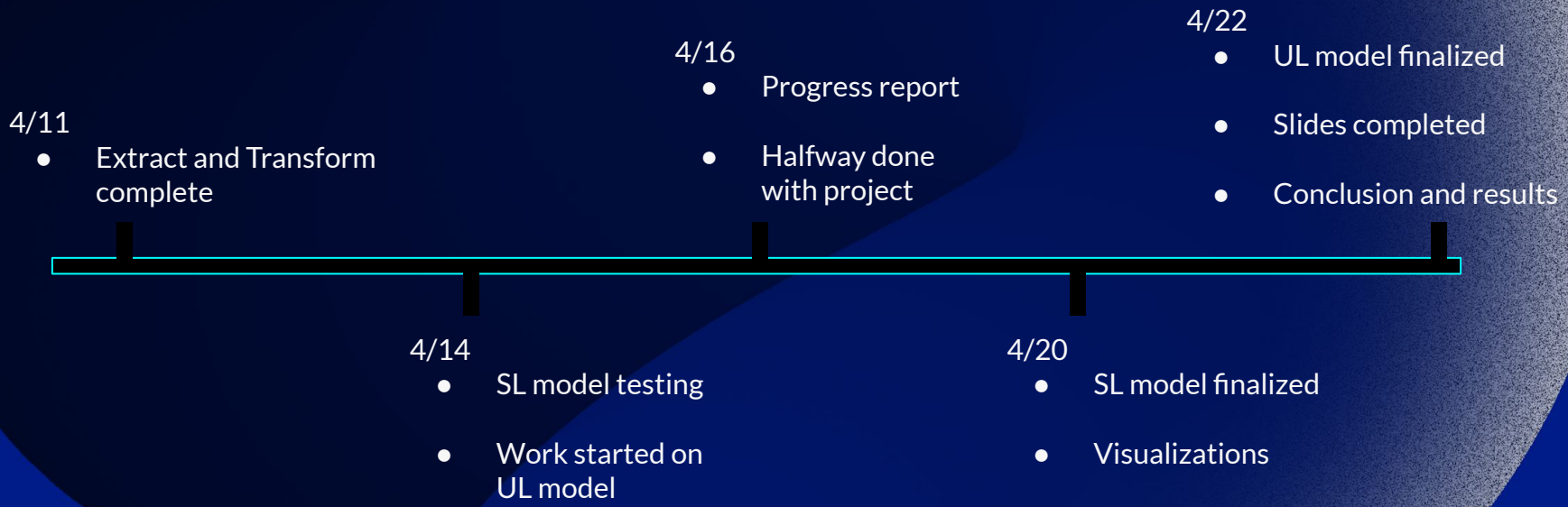
I310D Teclo Dataset

Group: NONAME001

Ethan Eckmann
Preston Cusick
Daniel Wendland
Rian Kahlon

Timeline


4/11
Extract and Transform complete
4/14
SL model testing
Work started on UL model
4/16
Progress report
Halfway done with project
4/20
SL model finalized
Visualizations
4/22
UL model finalized
Slides completed
Conclusion and results

Introduction

The problem:

As a telecommunications provider, Telco's number one priority is preventing "churn"

Churn - the rate at which customers discontinue paid services

Objective

Identify the main factors driving customer churn in the Telco dataset

Offer solutions to the Telco company that would help lower the rate of customer churn in the future

Data Overview

- Telco customer churn: IBM dataset from kaggle.com
- Some redundant or erroneous data fields
- Transformation needed before use in ML models
- Recorded any changes to the dataset

Data fields

Dropped columns:

CustomerID
Count
Country
State
City
Zip code
Lat Long
Latitude
Longitude
Phone Service
Internet Service
Contract
Churn Label
Churn Score
Churn Reason

Used columns:

Gender
Senior Citizen
Partner
Dependents
Tenure Months
Multiple Lines
Online Security
Online Backup
Device Protection
Tech Support
Streaming TV
Streaming Movies
Paperless Billing
Payment Method
Monthly Charges
Total Charges
Churn Value 1=yes/0=no
CLTV (higher =, more likely to default on loan)

Key:
Numerical
Boolean
String

Methodology

ML approaches:

- Supervised learning

Train ML model on the dataset and test F1 scores for different ML algorithms

- Unsupervised learning

Based on the predictions from SL model, categorize entries into groups based on risk

- Additional visualizations and interpretability

Why?

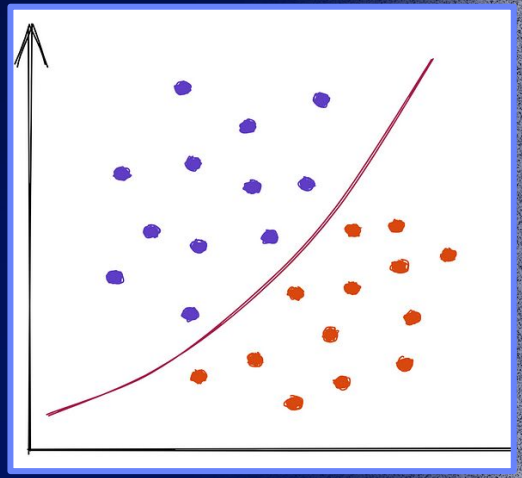
Supervised learning

1. Logistic Regression
2. Random Forest
3. XGBoost

Logistic Regression

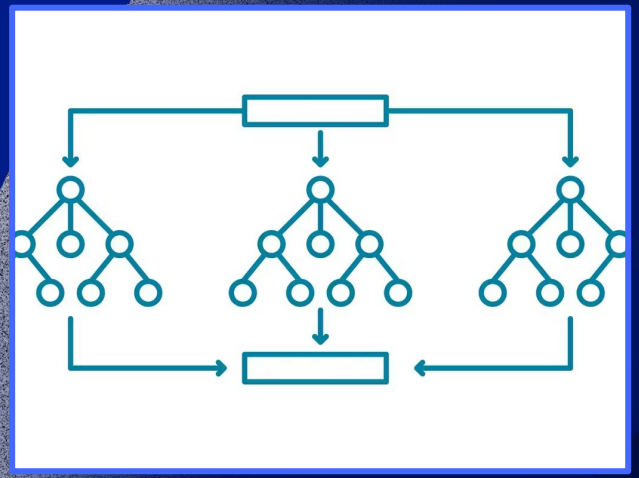
Metric	Score
Accuracy	79.30%
Precision	67.40%
Recall	52.30%
F1-score	58.90%

	Predicted No Churn	Predicted Churn
Actual No Churn	908	101
Actual Churn	191	209



Logistic Regression draws a linear line between churn and no-churn, predicting based on the weighted sum of features.

Random Forest



Random Forest uses many decision trees at the same time, combining their results through voting to improve accuracy.

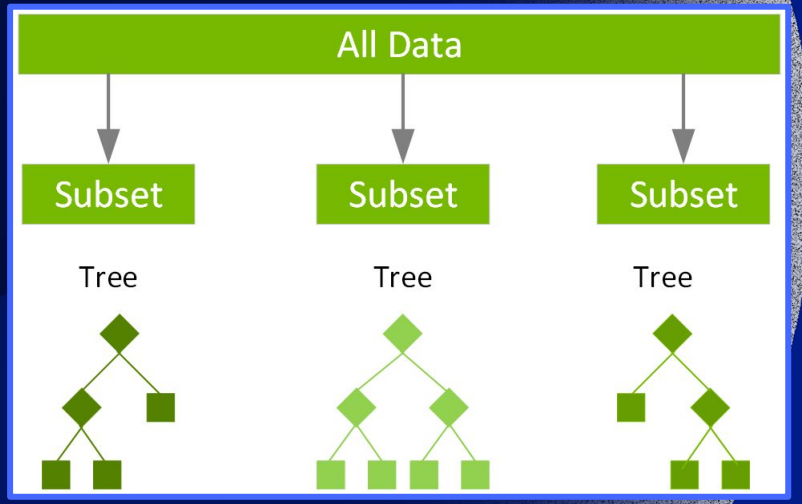
Metric	Score
Accuracy	79.50%
Precision	69.20%
Recall	50.00%
F1-score	58.10%

	Predicted No Churn	Predicted Churn
Actual No Churn	920	89
Actual Churn	200	200

XGBoost - Chosen model

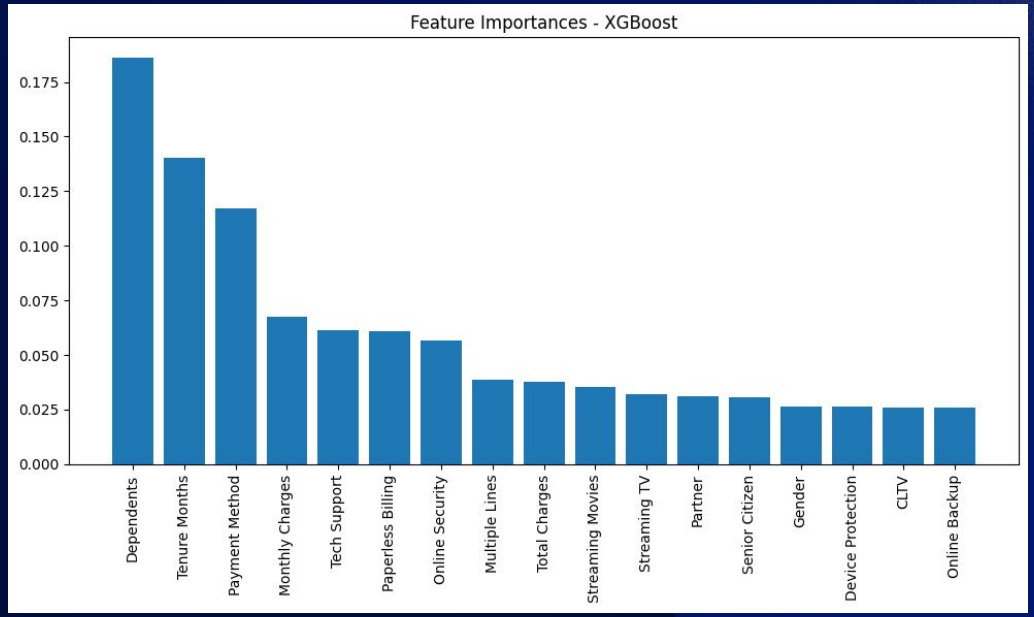
Metric	Score
Accuracy	79.90%
Precision	68.60%
Recall	54.00%
F1-score	60.40%

	Predicted No Churn	Predicted Churn
Actual No Churn	910	99
Actual Churn	184	216



XGBoost builds many small decision trees sequentially, each learning from the previous tree's mistake, making it more accurate for complex patterns.

Findings of Model



Top 3:

1. Dependents
2. Tenure Months
3. Payment Methods

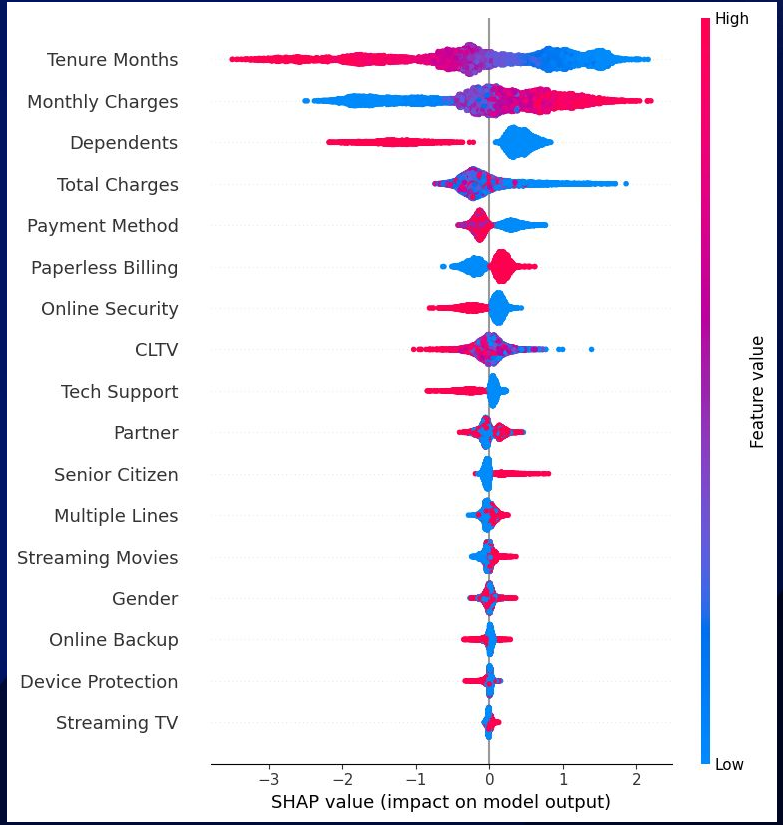
Understanding the Model

Payment method average impact on churn

- Electronic check = 0.3262
- Mailed check= -0.1733
- Auto-pay: = -0.1243

Right indicates positive push for churn
Left indicates negative push for churn

Red = high value
Blue = low value



Unsupervised learning

- Applied K-Means Clustering on churn probability scores predicted by the supervised ML model
- Focused on segmenting customers into risk groups using the “Probability” Field
- 4 clusters to distinguish between groups of risk
- Cleaned data by removing entries with missing Probability values

Clustering Results

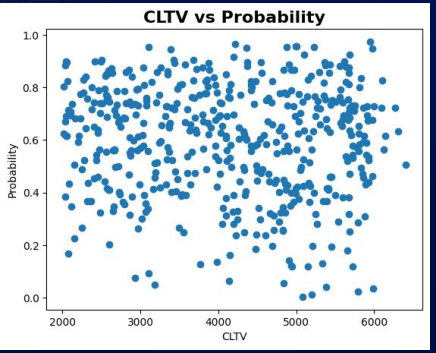
Cluster Distribution

- High risk - 595
- Moderate risk - 576
- Low risk - 438
- Very low risk - 260

- **31.83%**
- **30.82%**
- **23.43%**
- **13.91%**

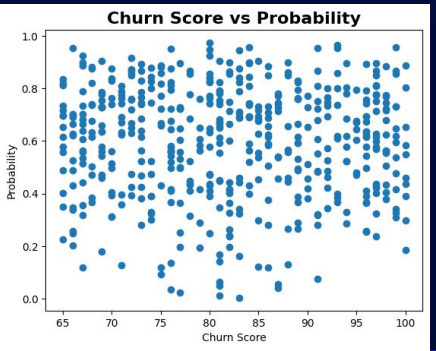
Average probability of churn for Telco customers: 0.5956878877768603

Visualizations



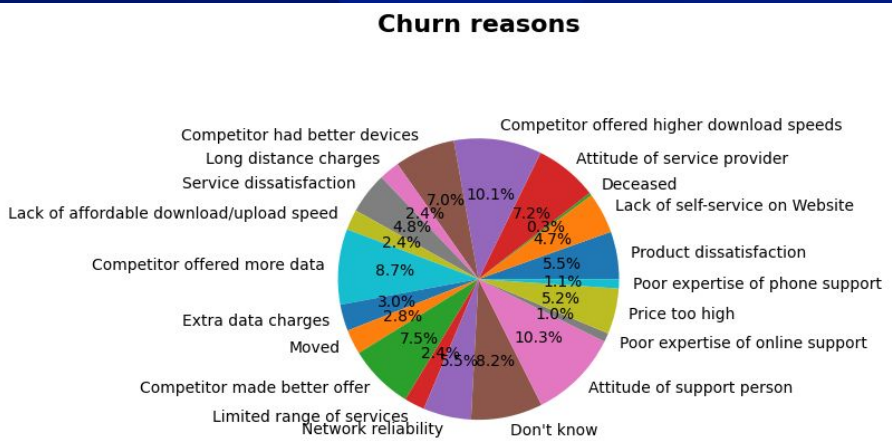
PCC = -0.10686437490942835
P-value = 3.654423200583537e-06

Not highly correlated



PCC = -0.020755943701178656
P-value = 0.3698171902387501

Not highly correlated



Insights/Results

1. Dependents

Customers with dependents at lower risk

2. Tenure Months

Total amount of months that the customer has been with the company

Higher tenure months lowers risk

3. Payment methods

Mailed check and Auto-pay: Lower probability

Electronic check: Higher probability

Conclusions

- Target customers with families
- Loyalty programs
- Encourage auto-payment

Limitations

1. Size of dataset
2. Original scope vs amount of factors
3. Time constraints

Q&A