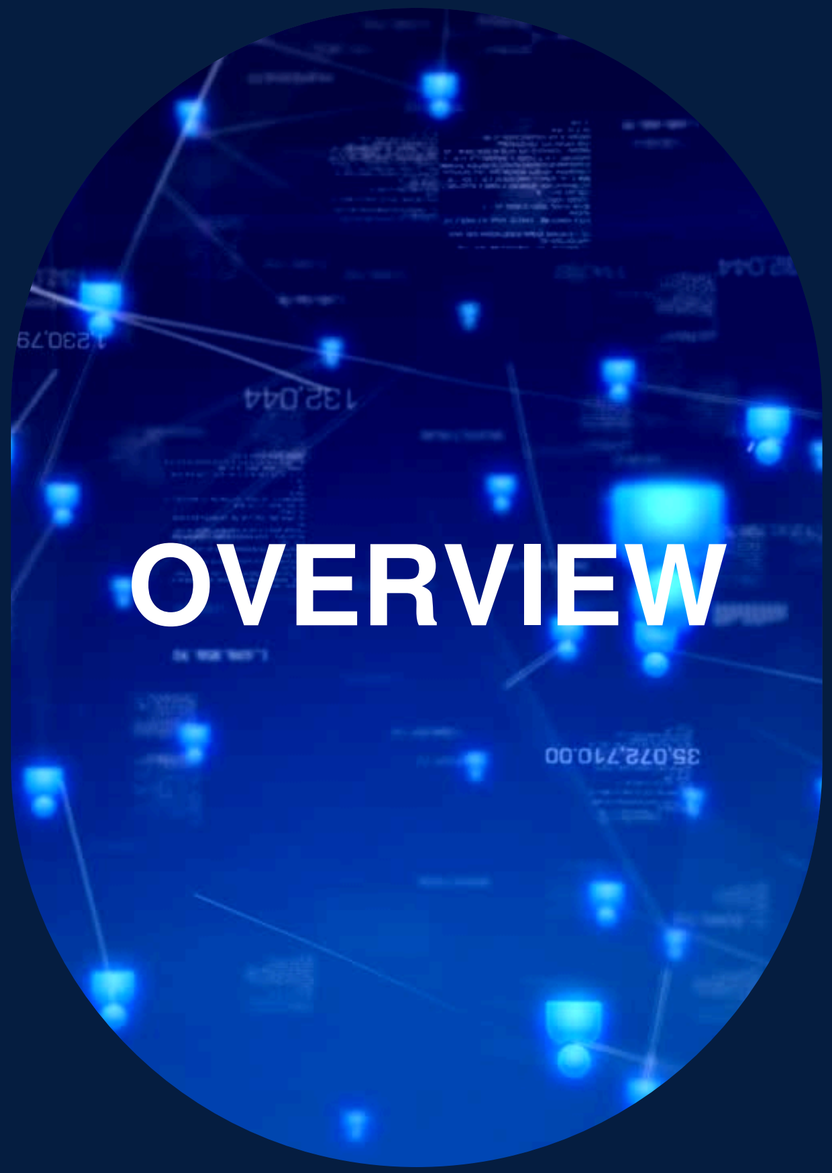
TELCO
CUSTOMER CHURN

Jonathan Beck



OVERVIEW

01 Problem Statement

02 Project Goals

03 Data Description

04 Initial Data Issues

05 Preprocessing

06 Data Visualization

07 Training & Validation Split

08 Baseline Model

09 Decision Tree Insights

10 Optimal Model

11 Decile Chart Insights

12 Final Analysis, Recommendation, & Strategy

01 PROBLEM STATEMENT

TELCO

Telecommunications Company.

Offer a variety of services to their customers such as Internet Service, Phone lines, and television streaming.

PROBLEM

Have noticed that roughly 26% of the 5000+ customers in database have defected to the competition.

Customer ending contract = lost revenue

02 PROJECT GOALS

01 Board of Directors at Telco are unsure what is most contributing to a customer deciding to leave Telco. They would like this issue to be identified so that the issue can be rectified accordingly.

02 Using the data on hand, they would like me to design a model that is able to predict propensity rates of customer churn.

03 As they want a better understanding of the factors contributing to customer churn, they would also like me to propose solutions to the marketing and sales department, so that the customer churn rate can be decreased going forward.

04 Ideally, they would like to see the customer churn rate brought down by 5-10%.

03 DATA DESCRIPTION

- **CustomerID**: The unique identifier (primary key) of the customer.
- **Gender**: Gender of Customer (Male or Female).
- **SeniorCitizen**: Whether or not the customer is identified as a senior citizen.
- **Partner**: Whether or not the customer has a partner.
- **Dependents**: Whether the customer has dependents or not.
- **Tenure**: Number of months the customer has stayed with the company.

- **PhoneService**: Whether the customer has subscribed to a phone service with the company.
- **MultipleLines**: Whether or not the customer has multiple phone lines with the company.
- **InternetService**: Whether the customer has subscribed to an internet service plan (Telco offers DSL and Fiber Optic)
- **OnlineSecurity**: Whether the customer has the Online Security add-on.
- **OnlineBackup**: Whether the customer subscribed to Online Backup.
- **DeviceProtection**: Whether the customer subscribed to device protection.
- **TechSupport**: Whether the customer subscribed to receive Tech Support.
- **StreamingTV**: Whether the customer subscribed to receive streaming television.
- **StreamingMovies**: Whether or not the customer subscribed to stream movies with their plan.
- **Contract**: The contract-term length of the customer (Month-to-Month, One year, Two Year).
- **PaperlessBilling**: Whether the customer has a paperless billing feature or not.
- **PaymentMethod**: How the customer pays (Electronic check, Mailed Check, Automatic Bank Transfer, or Credit Card).
- **MonthlyCharges**: The amount charged to the customer monthly.
- **TotalCharges**: The total amount charged to the customer.
- **Churn**: Whether the customer churned or not.

04 INITIAL DATA ISSUES

Numerous Features with a large amount of missing values.

Numerous features that would expect to be of "Integer" data type stored as object e.g., Senior Citizen having character values rather than binary digits.

Total Charges stored as an object rather than as floating point decimals.

```
                                Non-Nul
                                --------
 nnamed: 0                      5043 non-null
 customerID                     5043 non-null   ob
 gender                         5043 non-null   obj
3  SeniorCitizen                5043 non-null   objec
4  Partner                      5043 non-null   object
5  Dependents                   5043 non-null   object
6  tenure                       5043 non-null   int64
7  PhoneService                 5043 non-null   object
8  MultipleLines                4774 non-null   object
9  InternetService              5043 non-null   object
10 OnlineSecurity               4392 non-null   object
11 OnlineBackup                 4392 non-null   object
12 DeviceProtection             4392 non-null   object
13 TechSupport                  4392 non-null   object
14 StreamingTV                  4392 non-null   object
15 StreamingMovies              4392 non-null   object
16 Contract                     5043 non-null   object
17 PaperlessBilling             5043 non-null   object
18 PaymentMethod                5043 non-null   objec
9  MonthlyCharges               5043 non-null   floa
   TotalCharges                 5038 non-null   ob
   hurn                         5042 non-null
   loat64(1), int64(2), object
        866.9+ KB
```

05

PREPROCESSING

BOOLEAN VALUES

Replaced all "Binary Strings", such as "0", "No", "1", "True"" found across the binary categorical columns with appropriate boolean values.

TOTAL CHARGES

Converted the 'Total Charges' column into a FLOAT data type.
In doing this, I found that a small handful of the records contained empty strings in this column rather than a missing value or a 0. Handled this with K-Nearest Neighbors imputation.

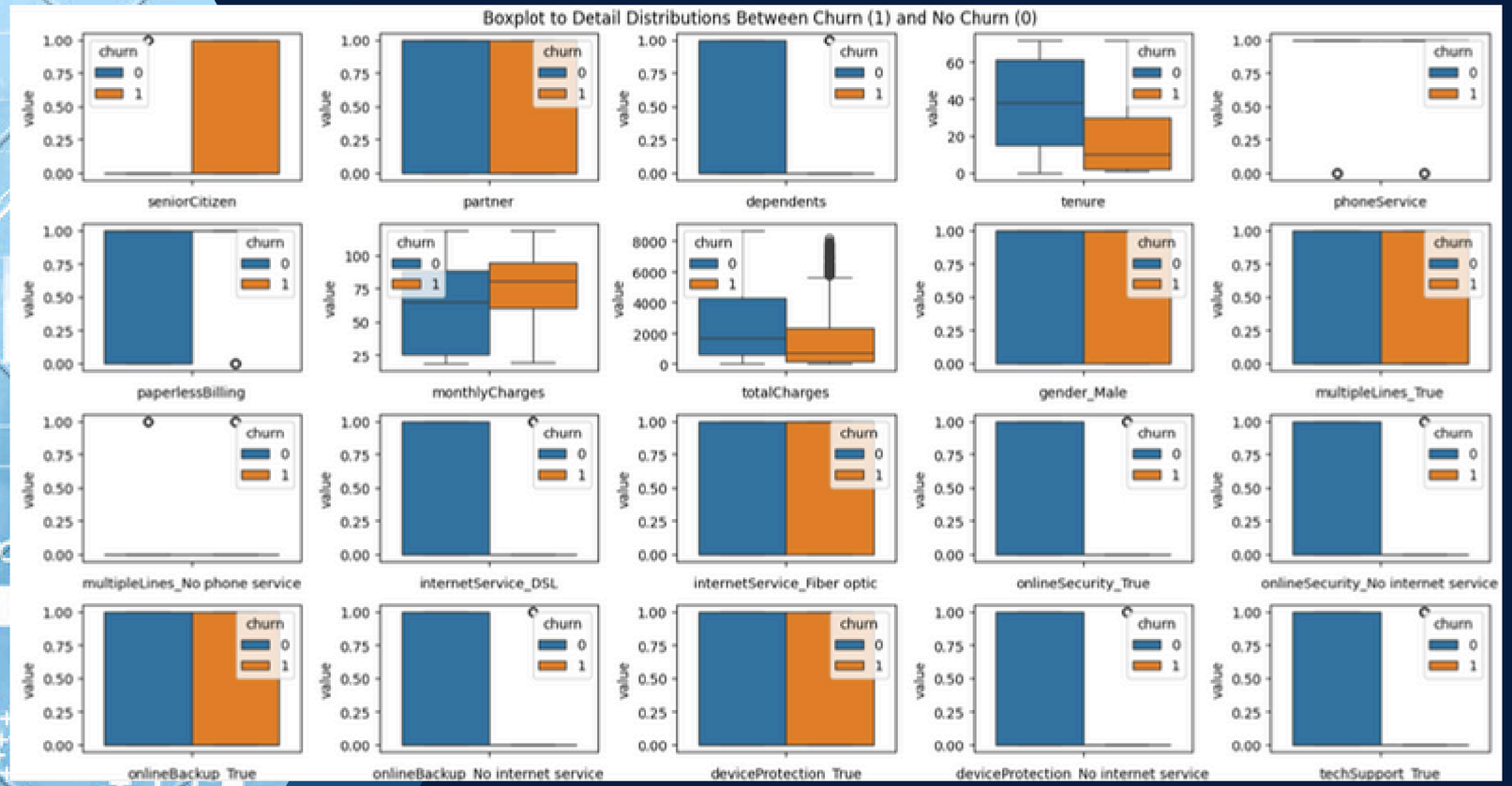
HANDLING MISSING VALUES

Found that, typically, a feature contained a null value if its "parent" feature contained a False value. E.g., if a customer didn't have "Phone Service" than a feature such as "Multiple Lines" either contained No Phone Service or a null value. For these instances, I replaced the Null with a "No __ Service" string value.

One record dropped. Missing response value.

DUMMY VARIABLES AND BINARY 0/1 TO REPRESENT "CHURN" VS "NO CHURN"

06 DATA VISUALIZATION



07 TRAINING & VALIDATION SPLIT

90%

Used for training. Dropped the "Partner" feature. 4537 customer records.

churn	count
0	3335
1	1202

Set the stratify parameter to "y" in splitting to ensure equivalent proportion of "churn" and "no churn" customers was maintained in both training and validation sets because imbalanced class distribution.

10%

Used for validation. 505 customer records

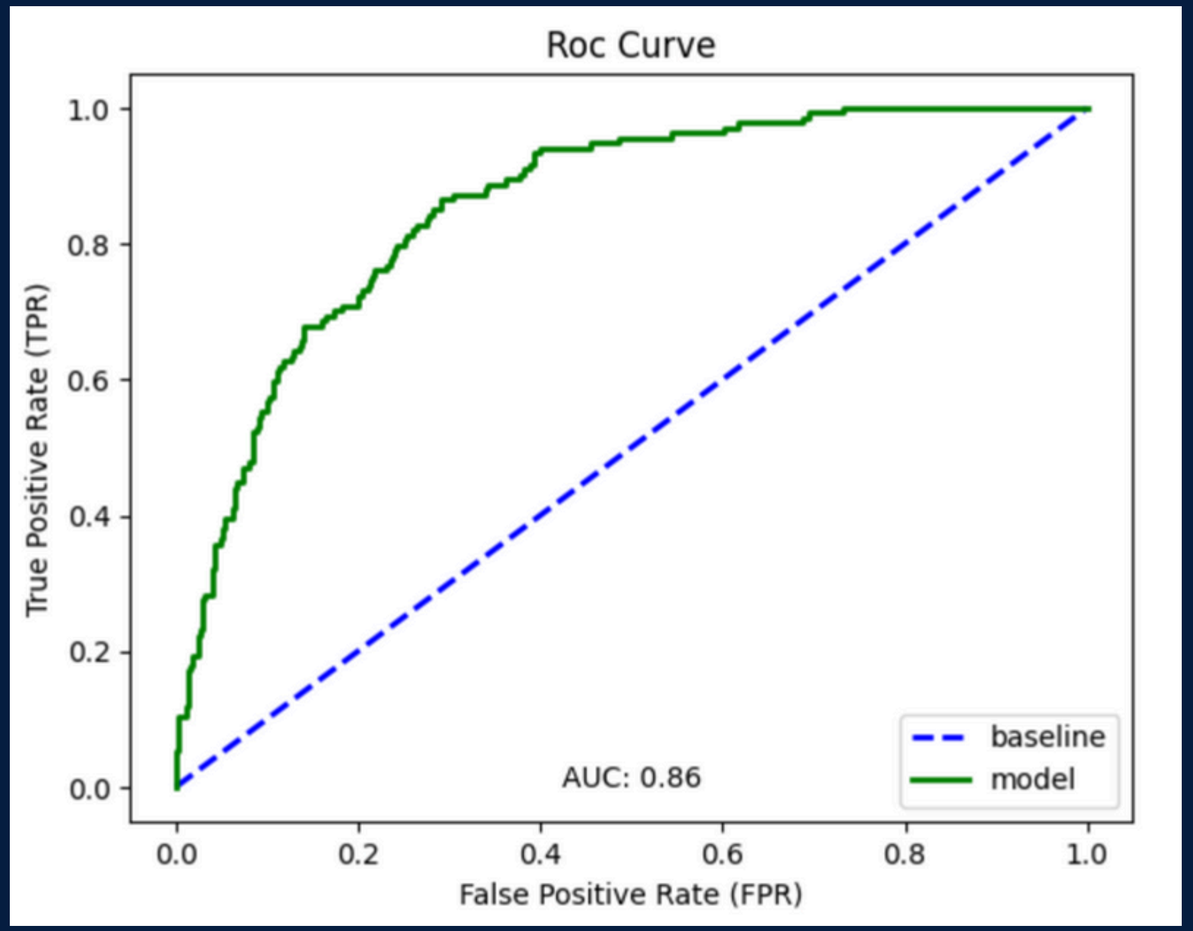
churn	count
0	371
1	134

08

BASELINE MODEL

LOGISTIC REGRESSION

- Max Iterations -> 1000
  - Maximum number of iterations solver runs before choosing beta coefficients that minimize loss function.
- Class Weights -> Balanced
  - Wrong predictions on the minority class/class of interest (Churn) carry more weight. Therefore, model doesn't become biased towards over populated class (No Churn)



Confusion Matrix (Accuracy 0.7644)

	Prediction	
Actual	no churn	churn
no churn	277	94
churn	25	109

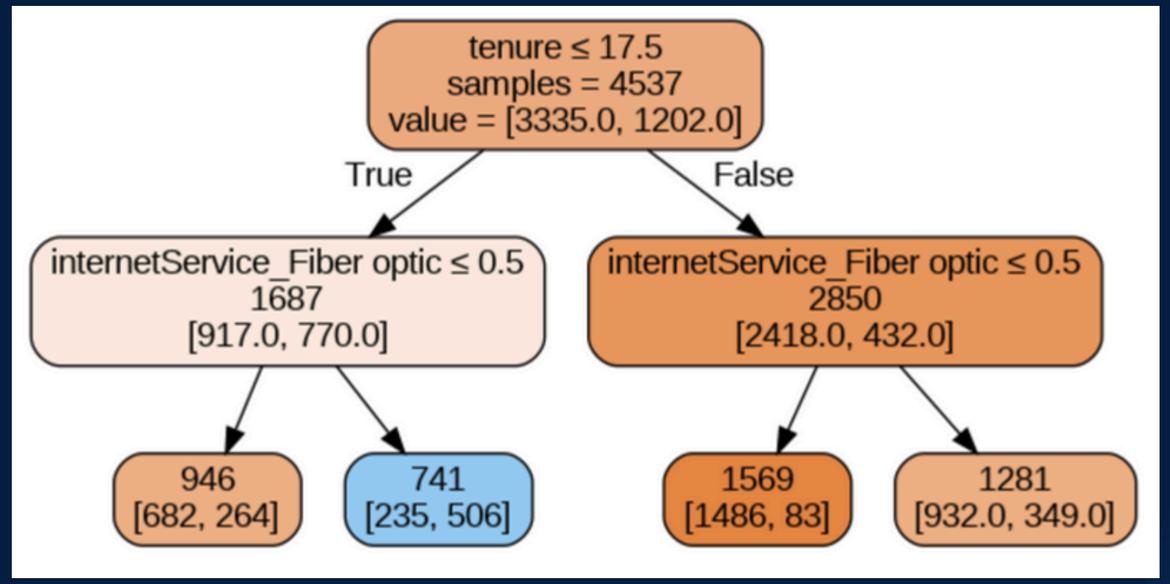
- **TPR**: 81.34%
- **FNR**: 18.66%

09

DECISION TREE INSIGHTS

Used a Decision Tree visualize most important features and their interactions.

- Although the Decision Tree algorithm didn't perform as well as my eventual best Logistic Regression Model, provided my first major insight.



Of the 741 customers in the training data who have been identified as 'doing business' with Telco for less or equal to 17.5 months *and* having subscribed to their Fiber Optic internet service package, 506 of them left TelCo. We also see that, of the 1,569 customers of TelCo's who have been with the company *longer* than 17.5 months and *don't* have the Fiber Optic internet service package, only 83 of them churned from TelCo. The other 1,486 customers have remained loyal customers.

Business insight: The Fiber Optic internet service package seems to be a *major* indicator of whether a customer remains with TelCo or doesn't. Having this package seems to greatly increase the likelihood that a customer will distance themselves from doing business with TelCo altogether. Moreover, it appears that customers who *do* have Fiber Optic IP from TelCo, but have been with the company longer than 17.5, does not in any way determine whether a customer leaves or not. This could be because these customers bought the service at a time when the service provided faster download speeds and was meeting customers expectations. Perhaps, the download and upload speed of the Fiber Optic service TelCo offers falls shorts of the expectations of new customers who have recently switched to TelCo, from another internet service provider. As such, this service needs to be properly analyzed and reconfigured before it is marketed towards TelCo's customers.

10 OPTIMAL MODEL

I decided that setting the cutoff threshold at 0.45 was optimal for my data mining task. The overall accuracy only decreased by roughly 2% when setting this threshold; moreover, and most importantly, the ***True Positive Rate*** increased by almost 6%, all the way up to 94%. In fact, the model only had issue correctly inferring 8 of the total 134 churn customers in the validation data.

LOGISTIC REGRESSION @ 0.45 Threshold

- Standard Scaling
  - Helped model converge on best solution and greatly improved efficiency of hyperparameter tuning using Grid Search Cross Validation
- Grid Search CV Settings
  - K-Fold Cross Validation -> 5
  - Scoring -> Recall
    - Maxize True Positive Rate. Naturally lowers False Negative Rate at expense of decreased Accuracy Rate
- Selected Hyperparameters
  - Class Weight -> Balanced
  - Penalty - > L2
    - Discourages large coefficients but does not drive any to 0.
  - C -> 5
    - Permits slightly larger coefficients.
  - Intercept -> False
  - Solver -> lbfgs
    - Loss Minimization function. Supports L2 regularization specifically.

Confusion Matrix (Accuracy 0.7188)

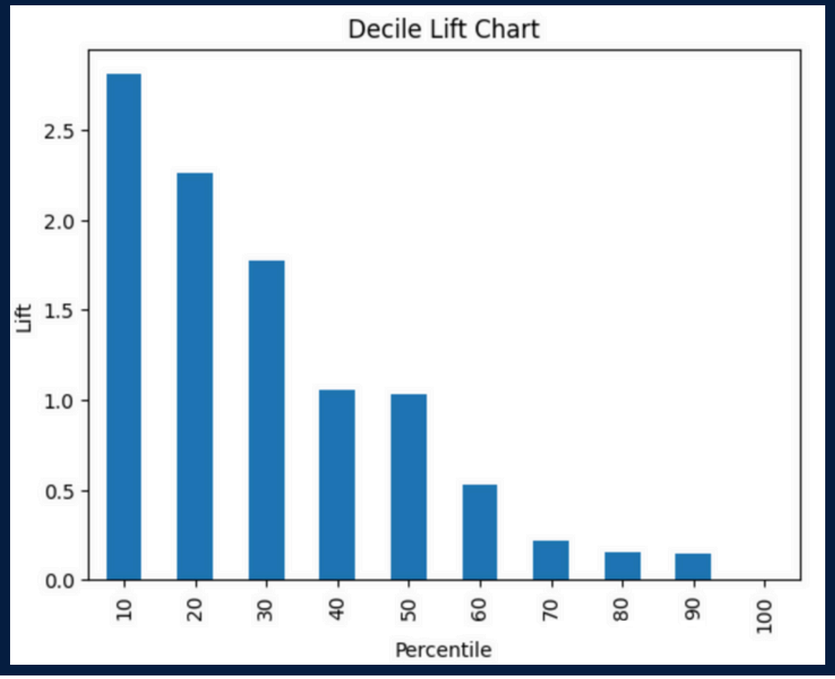
	Prediction	
Actual	no churn	churn
no churn	244	127
churn	15	119

Confusion Matrix (Accuracy 0.6931)

	Prediction	
Actual	no churn	churn
no churn	224	147
churn	8	126

DECILE CHART INSIGHTS

customerID	predicted	actual	prob_churn
7024-OHCCK	1	1	0.954586
4695-WJZUE	1	1	0.942162
8884-ADFVN	1	1	0.941698
2446-BEGGB	1	1	0.941602
1724-BQUHA	1	1	0.940507



The top 10% of customers that are ranked as 'most likely to churn' by our model yields over 2.5 more times as many actual churn customers compared to a random selection of 10% of customers from the validation data.

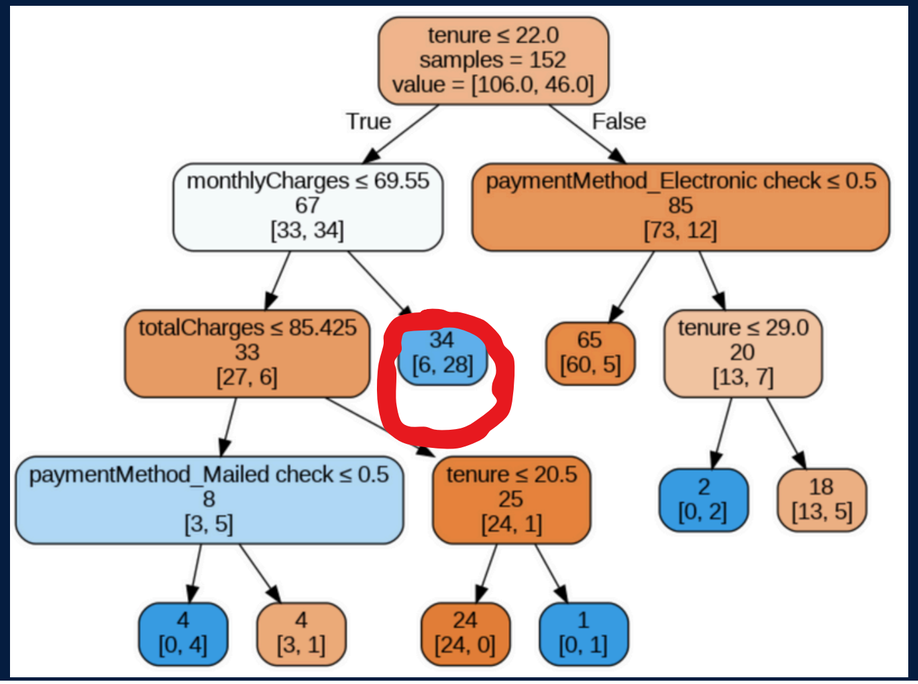
Given the insights provided by the above graph, I decided it would be worth analyzing more closely the top 30% of the customers in the validation data and take note of what specific characteristics they share and what is causing these 'churners' to leave TelCo, to see if the findings were agreeable with my earlier discoveries.

12

FINAL ANALYSIS

STRATEGY FOR ANALYSIS

RESULTING TREE



Decided to fit a Decision Tree to the top 30% of my validation data on hand.

- Methodology
  - Merged the original validation dataframe (consisting of 505 customers) to the ad-hoc 'predicted classes''dataframe I had created to construct the Decile Chart. Simple becaused customerID was kept as primary key in both.
  - Sorted records in descending order, based on "probability of churn".
  - Used the iloc method to only run the first 152 records through the tree.
  -

We see that the customers from this dataset who have been with TelCo for less than or equal to 22 months *and* have monthly charges *greater than* $69.55 are more likely to leave. In fact, slightly more than 82% of the 34 customers belonging to this group did in fact churn from TelCo.

However, it is noticeable that, on the validation data, the Tree did *not* use the "Fiber Optic" variable as a split. However, based on the earlier insight, I decided to analyze this specific node further and see how many of these customers had purchased the "Fiber Optic Service" from TelCo.

```
merged_df.iloc[:152].query("tenure <=22 & monthlyCharges>69.55 & predicted==1 & internetService_Fiber_optic==1")
```

churn	count
1	27
0	5

Business Analysis: We see that 32 of the 34 customers that belonged to that group *did in fact* have the Fiber Optic internet service from TelCo. We also see that, of those 32 customers, only 15% of them did not churn.

FINAL RECOMMENDATION & PROPOSED STRATEGY FOR TELCO

Recommendation & Strategy: The Fiber Optic service is not meeting customer expectations. I believe that TelCo is seeing a good number of new customers switching *to* TelCo and purchasing Fiber Optic internet service; however, the monthly rates they are seeing by subscribing to this service is not worth the perceived value that they are placing on its performance. The combination of high monthly charges and low value provided by the Fiber Optic internet package is a *primary* and undeniable factor leading to the high churn rate that TelCo is experiencing.

To prevent more churn, it is vital that TelCo lowers the monthly cost of this service and, while the cost is brought down, focuses on bringing the performance of the Fiber Optic service up to at least the performance and speed provided by its competitors. **Upon correcting the Fiber Optic issue, TelCo would see up to a 17% decrease in churn rate.**

APPENDIX

01

Additional Techniques Attempted

02

Performance on Test Data

01

ADDITIONAL TECHNIQUES ATTEMPTED

- RANDOM FOREST
- DECISION TREES
- GRADIENT BOOSTING
- K-NEAREST NEIGHBORS
- VOTING CLASSIFIER
- INTERACTION TERMS & POLYNOMIAL FEATURES
- OVERSAMPLING
- FEATURE SELECTION

02 PERFORMANCE ON TEST DATA

Why 'Test Data'? This data acts as a proxy for real-world data, to estimate how well the model will perform upon deployment in a practical operational setting. Specifically, in the given business context, as TelCo attracts new customers, they will be able to perpetually infer the probability that a customer will remain loyal customers or defect to the competition. This serves as a tremendous competitive advantage for TelCo, as they can monitor any given customers likelihood of churning, and, if necessary, take appropriate actions to keep these probabilities low across their customer base.

Results with a Cutoff Threshold of 0.45

```
Confusion Matrix (Accuracy 0.6533)

          Prediction
  Actual no churn    churn
no churn      822      647
   churn       47      486
```

Total Customers in Test Data: 2002

- **Accuracy Rate**: 65.33%
- **Misclassification Rate**: 34.67%
- **True Positive Rate**: 91.18%
- **False Negative Rate**: 8.82%

Implementation

This additional data was obtained from Kaggle. Upon importing the data into my python environment, I discovered that it contained 7043 customer records and 20 features. Moreover, the feature variables were identical to that of the 'TelCo Customer Data' that I had imported and used in training and validating my model. I removed the duplicate customer rows (that had already been used in building my model), and that left me with 2002 additional customer records to ensure the validity of my model.

THANK YOU