Predicting Telecom Industry Customer Churn

Improving Retention and Increasing Lifetime Value

Matt Wladyka

April 2020

Motivation

- Lower rates of churn imply longer tenure
- Tenure clearly has high positive correlation with customer lifetime value
- How best to structure product offerings to increase customer retention

Who Can Benefit?

Telecom Providers

Any business trying to retain customers!

Data Overview

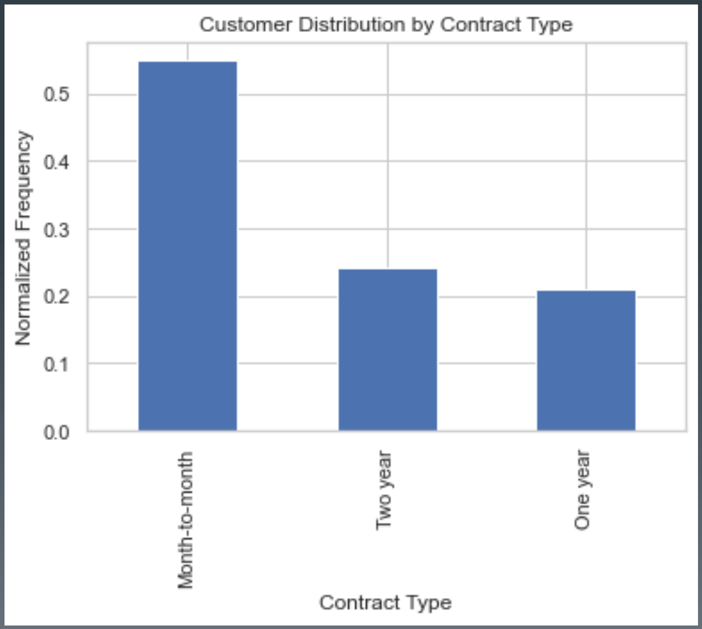
- 7,043 rows of unique customer data
- 20 unique columns (3 numerical, 17 categorical)
- Prediction variable of interest: Churn
  - Want to predict which customers are most likely to leave
- Right censored data (tenure no longer recorded after 72 months)
- Source: Kaggle Dataset
  - https://www.kaggle.com/blastchar/telco-customer-churn

Data Wrangling

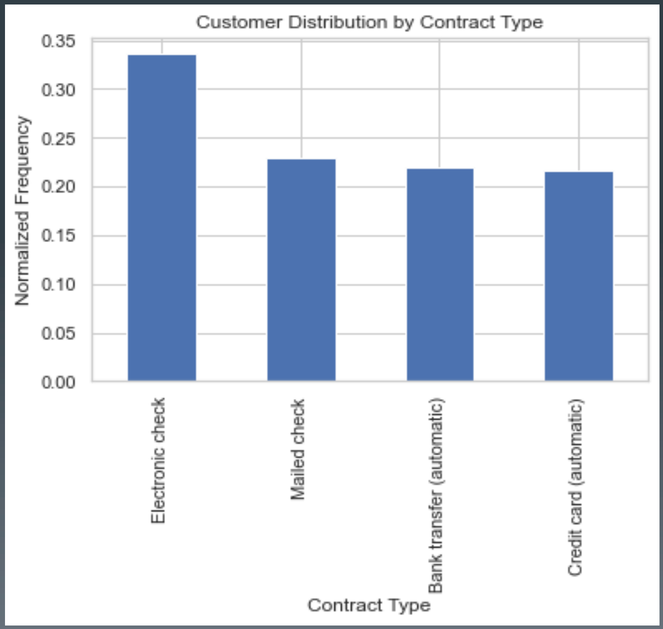
- Standard type conversions (categorical, int, float)
- Address missing values
  - Convert blank entries in CSV file to NaN
  - Only 11 rows with NaN values, removed these data points
- Remove Customer ID column, useless

Exploratory Data Analysis (EDA)

Preferred Contract: Month-to-Month

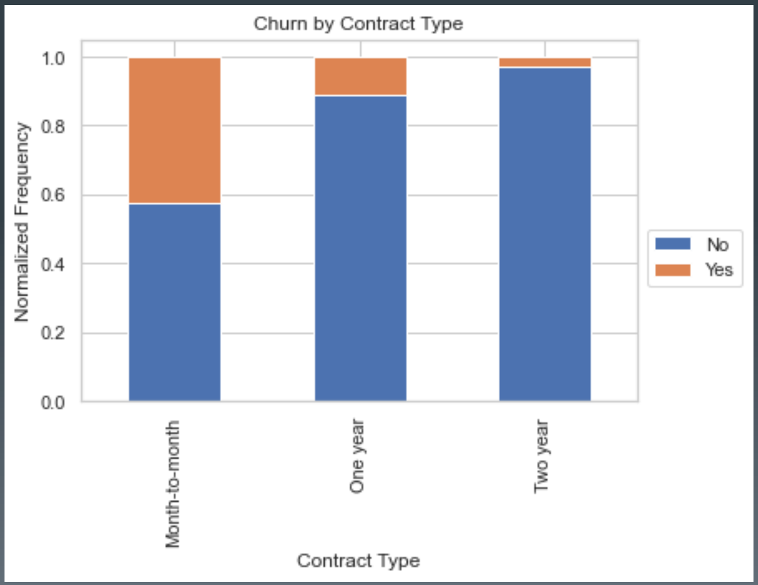


Preferred Payment: Electronic Check

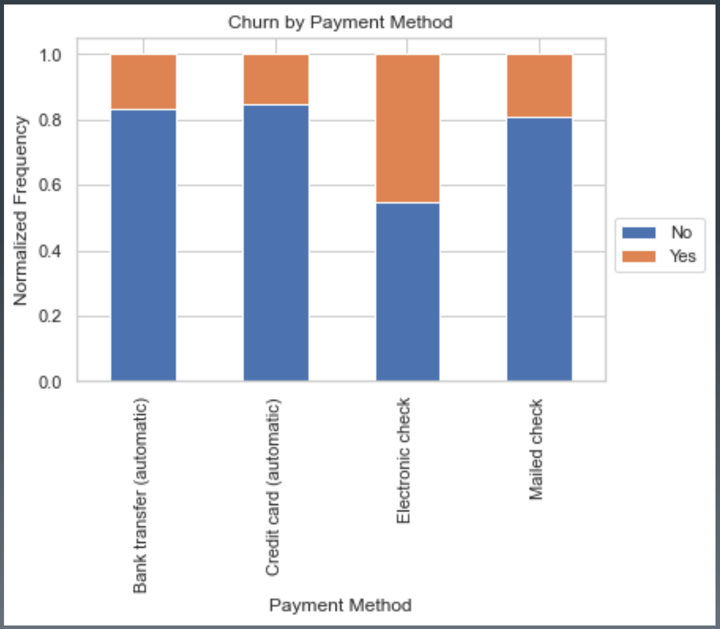


EDA

Highest Contract Type Churn: Month-to-Month

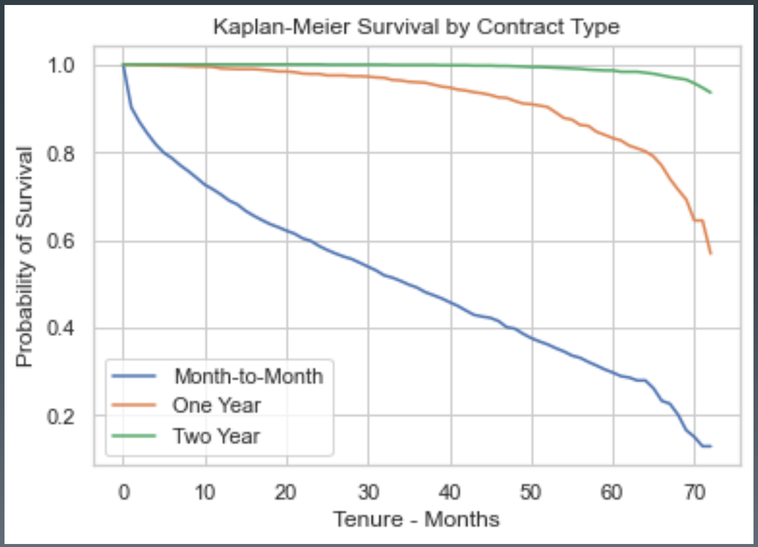


Highest Payment Method Churn: Electronic Check

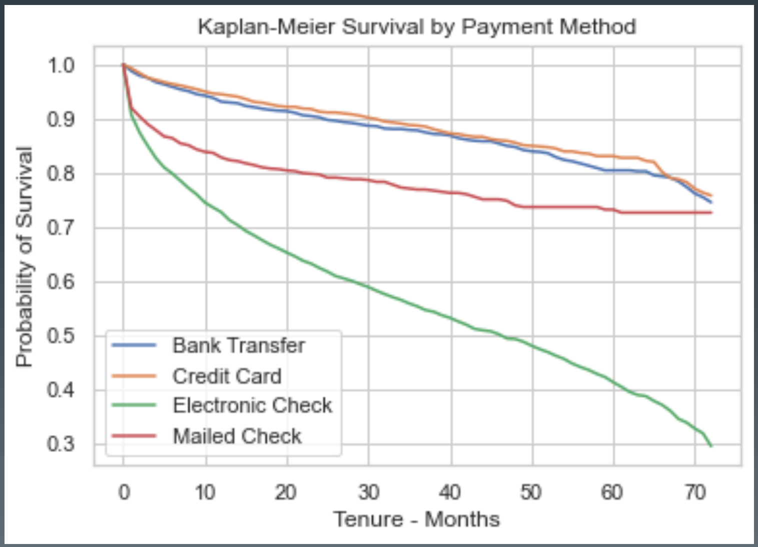


EDA

Month-to-Month: Survival Worse Across all Times



Electronic Payment: Survival Worse Over Time



EDA Takeaways

- Contract type with highest percentage of users: month-to-month
- Most used payment method: electronic check
- Thoughts for deterring customers from these methods:
  - Incentivize towards other plans/methods
    - Example: Show upcharge for electronic check or discount for other payment methods
  - Remove options: no choice for electronic check and potentially get rid of month-to-month option

EDA Takeaways

- Cost-benefit analysis of customers lost with options removed
- Potentially limited time offerings of no month-to-month or removed electronic check option for new customers
- Important to ensure profitability improves and customers are kept
- EDA findings alone not enough
- Further solidify our analysis with in-depth analysis of relevant machine learning models

Machine Learning: Modeling Overview

- Binary Classification Problem (Churn or not churned)
- Data pre-processing
- Modeling pipeline
- Model selection & fitting
  - Hyperparameter tuning
- Model Comparisons
- Conclusions

Data Pre-Processing

- Encode Labels
- Split Data: 70% training, 30% held out for testing
- Scale Data (appropriate models only)
- Cross-validation: 5-fold, repeated twice, stratified

Modeling Pipeline

- Impute data for missing values
- Scale values as required by model type/performance
- Tune hyperparameters
- Analyze numerical and graphical prediction results

Model Selection

- k-Nearest Neighbors (kNN)
- Logistic Regression
- Random Forest
- Support Vector Machine (SVM)
- AdaBoost
- Voting Classifier

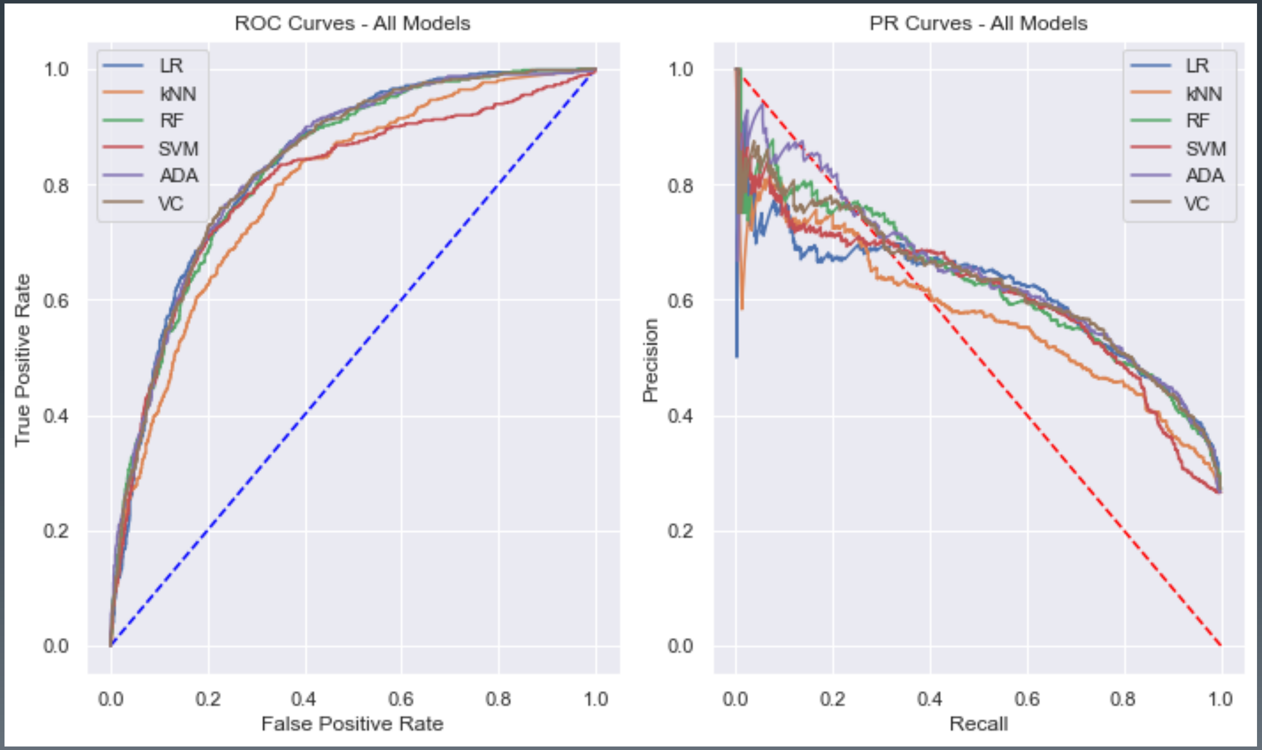
Metrics for Comparison

- ROC AUC – Area under ROC curve (compares tradeoff of true positive rate (TPR) and false positive rate (FPR) based on classification threshold)
- PR AUC – Area under precision recall curve
- Precision – Given a positive prediction, what is probability it is actually a positive result
- Recall – Given a random positive result, what is the probability it is correctly predictive as positive
- F1-score – Harmonic mean of either precision or recall

Model Comparison

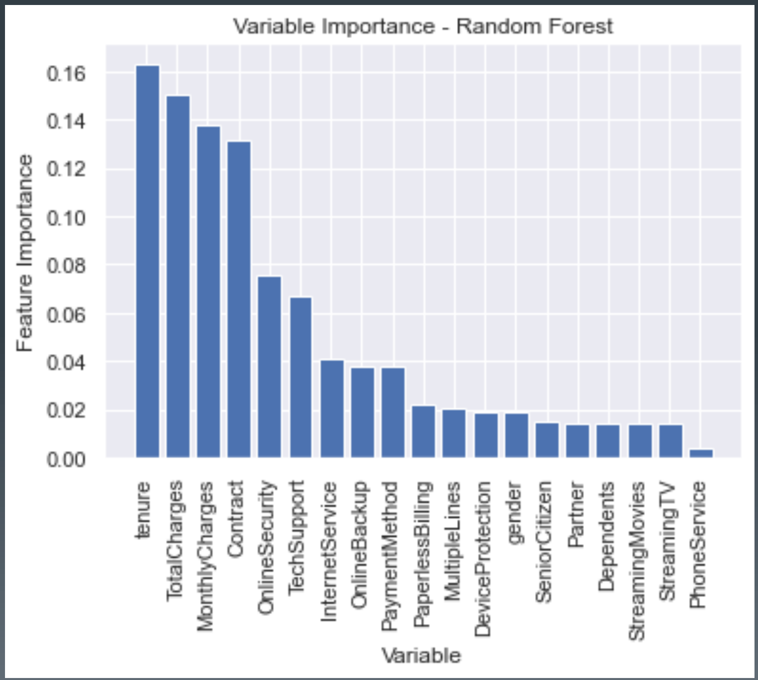
Model	Class	Precision	Recall	F1-score	ROC AUC	PR AUC	Accuracy
kNN	0	0.81	0.89	0.85	0.79	0.57	0.77
	1	0.59	0.42	0.49			
Logistic Regression	0	0.84	0.89	0.87	0.83	0.61	0.80
	1	0.64	0.55	0.59			
Random Forest	0	0.83	0.91	0.87	0.83	0.63	0.79
	1	0.65	0.49	0.56			
SVM	0	0.83	0.91	0.86	0.80	0.60	0.79
	1	0.64	0.47	0.54			
AdaBoost	0	0.84	0.89	0.86	0.84	0.64	0.79
	1	0.64	0.51	0.57			
Voting Classifier	0	0.82	0.91	0.87	0.83	0.63	0.79
	1	0.66	0.46	0.54			

Model Comparison: ROC and PR Curves



Random Forest: Feature Extraction

- Random Forest models allow for easy feature extraction
- Could be used for dimensionality reduction
- Confirms initial EDA findings: contract type and payment method important



Model Conclusions

- AdaBoost provides best overall results
- Logistic Regression strong performer for predicting minority class (customers who churn)
- Voting Classifier struggled relatively minority class
- Random Forest useful for feature importance analysis

Modeling Applications

- All models discussed predict classifications on customer churn and associated probabilities
- These are just snapshots of data today
- More useful to tell telecom client likelihood of churning at all future points in time
  - Will review one application: Random Survival Forest

Random Survival Forest

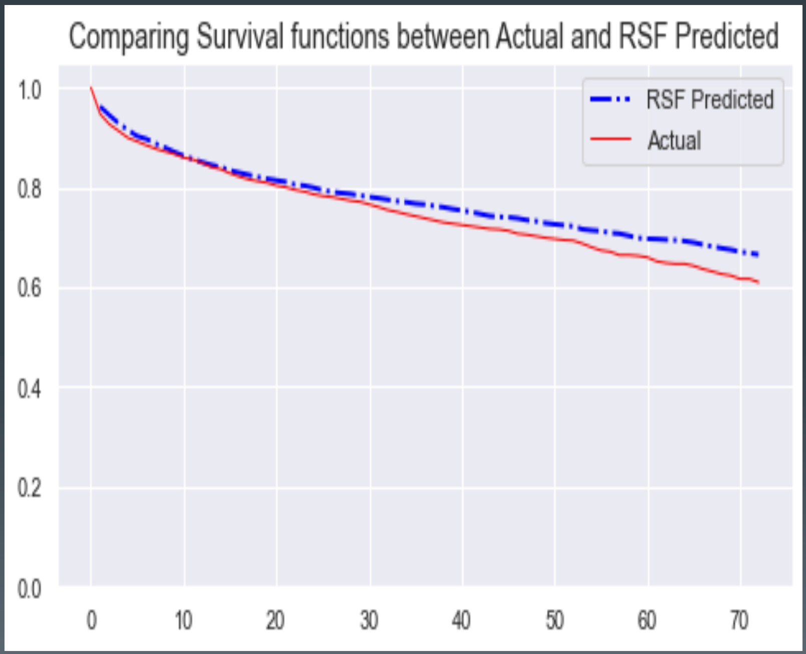
- Well equipped to handle censored data
- Dataset in this study is right censored
- Further extension of random forest model used in survival analysis

Random Survival Forest: Setup

- Again use split dataset (70% training, 30% test)
- Use same optimized hyperparameter values from random forest model
- Goal: generate survival curve for customer churn prediction at all points in time

Random Survival Forest: Results

- Concordance index (C-index): 0.88
  - Generalization of ROC AUC score for censored data
  - Best score of 1, random predictions score 0.5
- Integrated Brier score: 0.08
  - Best score of 0, random predictions score 0.25



Random Survival Forest: Conclusion

- Model is visually quite accurate at predicting survival curve
- C-index of 0.88 and IBS of 0.08 support this finding
- Useful in real-world applications for client
  - Ex: Customer has this churn profile, what is their probability of churning today and what does their survival curve look like?
  - Target marketing dollars accordingly

Future Work

- Feature Selection
- More models to compare with RSF (i.e. CSF, Extra Survival Trees)
- Deal with slight class imbalance: upsampling, downsampling, SMOTE
- Collaborate with telecom partner to score models set to optimize on client request (right now set to accuracy)

Conclusions

- AdaBoost model most likely used in practice
- Could develop risk ranking system for clients
- Coordinate with marketing department to target clients with higher risk of churning
- Prioritize based on churn risk today
- Create retention pipeline utilizing survival curve for future efforts

Contact Information

Matthew Wladyka, M.S. Mathematical Finance

Email: matt.wladyka@gmail.com

LinkedIn: https://www.linkedin.com/in/matt-wladyka-0364691a/

GitHub: https://github.com/wladykam