TelCo Customer Churn Prediction

INSY-695-076 Enterprise Analytics Final Project Report

Devanshu Khurma* Jiajun Huang† Charlie Cai‡ Evelyn Sun§

February 24, 2020

*260894480
†260629217
‡260926881
§260915480

Contents

1 Introduction

The core objective is predicting customer churn behavior for the TelCo Company. TelCo Company currently has no data-driven and quantitatively-based prediction model. Our team will develop machine learning models, using a 7043-observation (customer) with 20 predictors dataset. The model will be a supervised, binary classification model, since we have a clear binary target variable, churn or no churn. To achieve the optimal outcome, our team will both manually build machine learning models in Python and develop an Auto Machine Learning Model in H20.ai. The core performance will be measured in terms of the accuracy and F1 Score.

After the optimal model was chosen and picked, it will be run off-line periodically (at most once a day). In other words, the prediction model will build on a batch processing system using Apache Spark, which takes a large amount of input data, runs the pickled prediction model to process it, and produces the prediction outcome.

Our model will be used by the marketing and consumer retention team so that they could intervene early by either offering targeted online offers and coupons, or having representatives phoning those consumers who are predicted to churn. Also, the model explain-ability report will be used by middle-level managers to gain insights on the factor that both positively and negatively contribute to the churn rate so they could address such factors accordingly.

1.1 Project Hypothesis

The goal of our project is to predict (i.e., classifying) whether a current customer will churn from TelCo or not.

H_0(Null Hypothesis): The accuracy score is less or equal to 0.7 and no predicting power can be generated from the model.

H_α(Alternative Hypothesis): The accuracy score is greater than 0.7 and some predicting power can be generated from the model.

Types of Errors: Both Type I and Type II errors are expected to be made by the model. The goal of the model is to minimize Type II errors (i.e. reducing false negative rate.)

1.2 Core Actions

- Trained multiple Python-based Machine Learning models to predict customer churn for the TelCo Company
- Programmed a H20.ai-based Auto Machine Learning model and compared the AutoML model performance (in terms of f1 score and accuracy) against manually trained models
- Tested causal inference on following predictors: gender, SeniorCitizen, Partner, Dependents, PhoneService, and PaperlessBilling by applying Microsoft DoWhy package realized in Python
- Assembled a model interpretability and explainability analysis using SHAP package on XGBoost Classification Model
- Composed a Machine Learning Bias Report based on the SHAR analysis to suggest further managerial actions
- Lessons learned and next steps

2 Data Description

2.1 Data Acquisition

We used the data from Kaggle's Telco Customer Churn prediction challenge. The data has 7043 profiles of customers along with labels representing whether they churned or not.

The original dataset was 955 kb in size. We created a folder for the dataset after every stage of processing. We downloaded the data and stored in an csv file.

In addition, to ensure sensitive information is deleted or protected we removed the customer ID column during pre-processing data(e.g. anonymized). After that, 2e looked at the types of data and dealt with all the categorical variable and made sure that there was no data snooping.

2.2 Data Cleaning

In the data cleaning part, we looked at the number of missing values and surprisingly, found none. Then, we started of looking at the data types and counting features of each type, it turns out that 18/21 columns were categorical, 2 were integers and 1 was float.

Next, we looked at the number of different values there can exist in each of the categorical feature columns. It is noticed an issue here because TotalCharges column representing the total amount charged to the customer was a string and not a float like it should be, so we converted it to a float and realized that there were now some missing values for float which we replaced by 0 assuming the person was not charged anything so far.

In the last step, we dummified the categorical variables and moved on to exploration.

2.3 Data Exploration

In this step, we created a Jupyter notebook to keep a record of our data exploration.

first, we looked at summary statistics of the numerical columns to find the mean and standard deviation of customers who churned and did not churn. There was not much difference between the mean values of both classes' mean value for Monthly Charges. The variance in Monthly Charges for those who churned was about 33.33% the value of the mean while for those who did not it was about 50% of the mean value indicating the loyal customers showed higher diversity in the amount they paid monthly.

We then looked at the top features showing the highest the lowest correlation with the outcome variables. Monthly charges were positively correlated while Total Charges and Tenure were negatively correlated with Tenure most negatively correlated with the customer

churning as expected. This indicated that the longer the person stays the less likely he is to ever churn.

To better understand the correlation, we visualized the correlation matrix, and followed this by experimenting with various models to predict the churn.

2.4 Data Preparation

- Dealt with missing data in TotalCharges column
- Encoding of categorical variables
- Feature subset selection
- Feature scaling
- After the preparation, there were no missing values and all categorical variables were dummified and all numerical variables are normalized by using an sklearn-pipeline.

3 Churn Prediction Modeling

3.1 Python-based Model

We experimented with various different classifier algorithms including random forests, decision trees, logistic regression, K Nearest Neighbors and even XG Boost.

The performance from Random Forests and XG Boost was most promising so we used grid search CV to fine tune these 2 models.

We got the best performance from the random forest with accuracy of 84.9% and an F1 score of 0.56

3.2 H20 Auto-ML Model

H2O's AutoML can be used for automating the machine learning workflow, which includes automatic training and tuning of many models within a user-specified time-limit. It has made it easy for non-experts to experiment with machine learning to set a benchmark.

We have used H2O package to build models to compare the results with our Python-based model.

Due to the limited available algorithms at this stage of H2O, we were only able to perform Random Forest and Gradient Boosting to compare with our Python-based model. However, we also explored deep learning, and asked H2O to self-select 10 best performing models.

Table 1: H2O AutoML Model Results

Model	Accuracy Score	F1 Score
Random Forest	0.8272	0.630814
Gradient Boosting	0.83457	0.627168
Neutral Network	0.82574	0.619318

Table 2: 10 Best Performing Models Selected by H2O AutoML

model_id	auc	mse
StackedEnsemble_BestOfFamily_AutoML_20200216_115946	0.850544	0.133636
StackedEnsemble_AllModels_AutoML_20200216_115946	0.850509	0.133678
GBM_5_AutoML_20200216_115946	0.848467	0.13321
GLM_1_AutoML_20200216_115946	0.847801	0.133984
GBM_grid__1_AutoML_20200216_115946_model_1	0.842607	0.148619
GBM_1_AutoML_20200216_115946	0.841779	0.136873
GBM_2_AutoML_20200216_115946	0.841048	0.136614
GBM_3_AutoML_20200216_115946	0.838457	0.138417
DeepLearning_1_AutoML_20200216_115946	0.83469	0.13864
GBM_4_AutoML_20200216_115946	0.832629	0.140648
XRT_1_AutoML_20200216_115946	0.832395	0.140071
DRF_1_AutoML_20200216_115946	0.828687	0.141467

As we can see from Table 1 and Table 2, H2O AutoML has produced high performance models with accuracy scores are all more than 80%, in this case, we REJECT H_0, which means that models built by H2O AutoML have prediction po

3.3 Comparison of Python-Basedwer. Models and H2O AutoML

Figure 1: Model Result Comparison

From Figure 1, we can see that python based Tunned Random Forest delivers the highest accuracy, 0.85, but for Gradient Boosting and XG Boost, H2O AutoML actually produces better result.

4 Further Discussion and Interpretation

4.1 Casual Inference

We tested causal inference on following predictors: Gender, Senior Citizen, Partner, Dependents, PhoneService, and Paperless Billing by applying Microsoft DoWhy package realized in Python. Results are shown in the following Table 3.

We found that only Contract Month-to-month and Paperless Billing have causal relationship with the target. However, we cannot draw conclusion lightly based solely on the result. Interpretability and explainability report is needed to further investigate the relationship.

Table 3: Casual Inference Report

Variables	P-Value	Causal Estimate
Contract Month-to-month	< 0.001	0.05997
Paperless Billing	< 0.001	0.04491
Gender	0.358	-0.00329
Senior Citizen	0.044	0.0442
Partner	0.464	-0.00122
Dependents	0.065	-0.02092
Phone Service	0.424	0.00988

4.2 Churn Model Interpretability and Explainability

To access the XGBoost Classification Model interpretability and explainability, we used the SHAR package to visualize the predictors' effect on the target variable, churn. SHAP (SHapley Additive exPlanations) is a game theoretic approach to explain the output of any machine learning model. The reason we choose the XGBoost Classification Model to analyze instead of the Random Forest Model because the Random Forest Model takes significantly longer compared to the XGBoost Model, and our team's laptops are unable to provide the results.

We first can to visualize the first prediction's explanation in Figure 2.

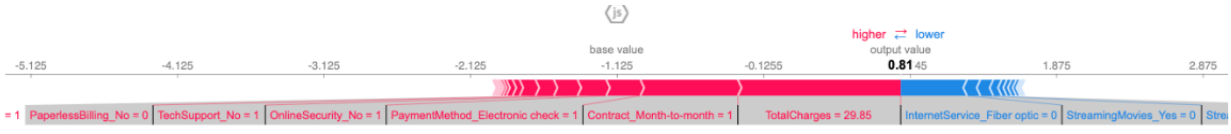

Figure 2

Next, we summarize the effects of all the features. The plot below sorts features by the sum of SHAP value, and uses SHAP values to show the distribution of the impacts each feature has on the model output. A higher feature value is colored in red and a lower feature value is colored in blue.

Here are some sample interpretations"

- If a consumer is on a month to month contract (red-colored), it is more likely to churn

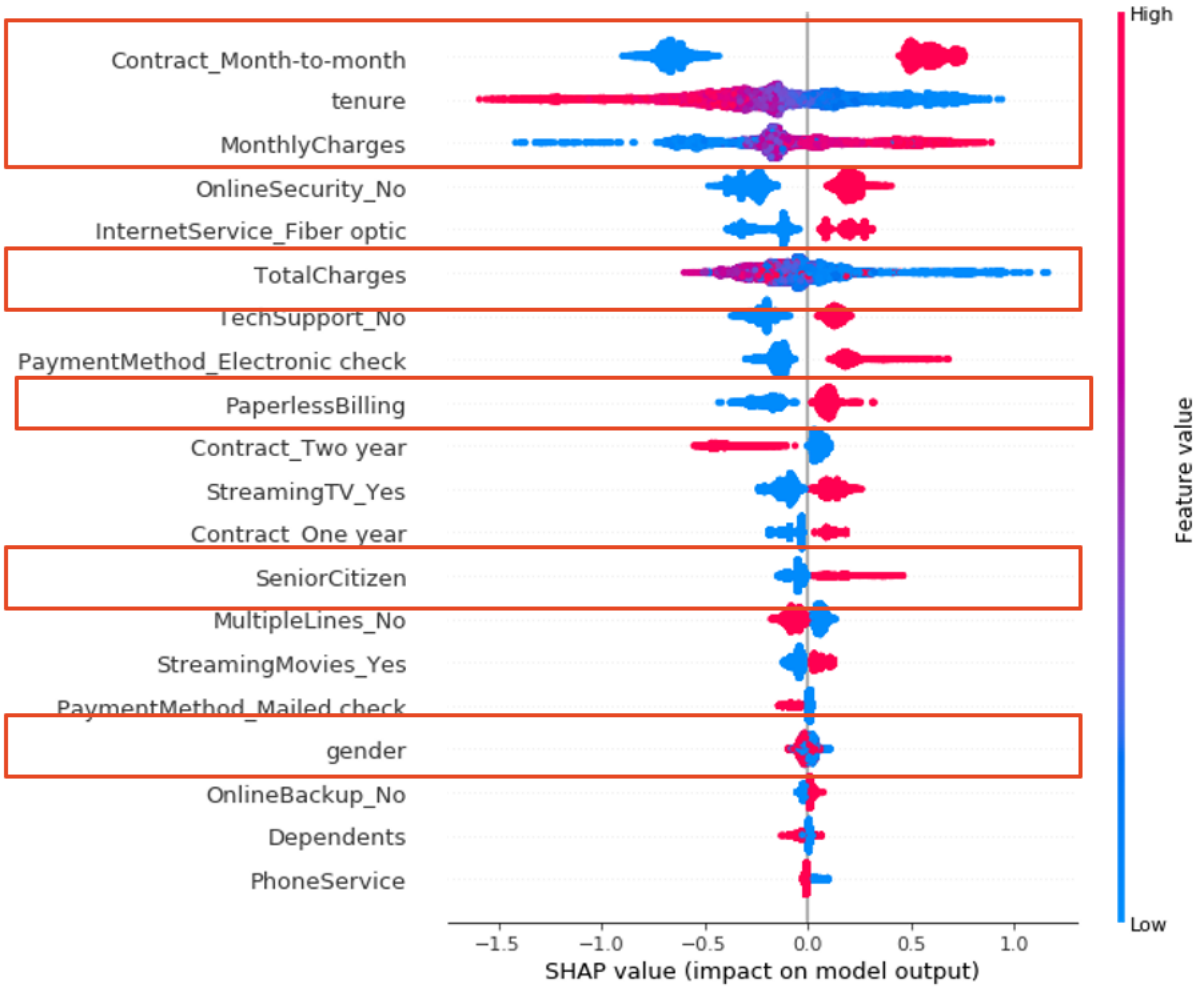
Contract_Month-to-month
tenure
MonthlyCharges
OnlineSecurity_No
InternetService_Fiber optic
TotalCharges
TechSupport_No
PaymentMethod_Electronic check
PaperlessBilling
Contract_Two year
StreamingTV_Yes
Contract_One year
SeniorCitizen
MultipleLines_No
StreamingMovies_Yes
PaymentMethod_Mailed check
gender
OnlineBackup_No
Dependents
PhoneService
−1.5
−1.0
−0.5
0.0
0.5
1.0
SHAP value (impact on model output)
High
Feature value
Low

Figure 3

- The longer a consumer is with the company (higher red-colored for tenure), it is less likely to churn
- If a consumer's monthly charges are high, it is more likely to churn
- On the other hand, if the total charges are high, it is less likely to churn. This could imply that for high spenders at TelCo, these consumers are less price sensitive
- If a consumer is a senior citizen, it is more lily to churn

We can also just take the mean absolute value of the SHAP values for each feature to get a standard bar plot (produces stacked bars for multi-class outputs):

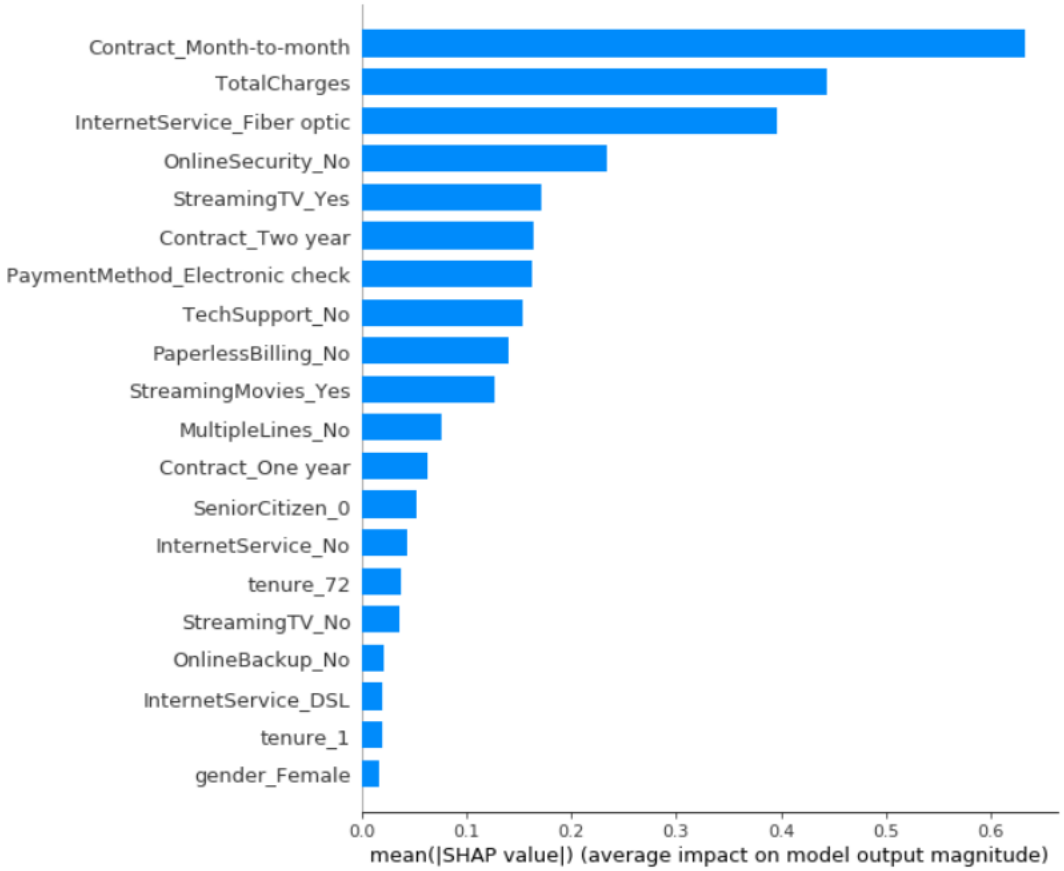

Figure 4

4.3 Threats to Validity

There are four primary considerations for data quality and threats to validity.

First, the data consist of only 7000 observations. The results we have are acceptable but are not outstanding. To achieve more desirable results, we recommend the TelCo to add more observations to train the model.

Secondly, the data provided is a snapshot of the past data at a certain point in time. However, as the market and competitors change, the customer churn will change as well. Hence we recommended the data analyst team the TelCo continuously train the model.

Thirdly, the model does not consider external factors, such as macro-economic factors.

Lastly, regarding the depth of the data, the data does not contain customer satisfaction levels or user engagement factors. We recommend the TelCo to explore more variables to include in the dataset.

5 Conclusion

5.1 Model Bias

There are four predictors used in the model that may be considered as potential biased or discriminatory, namely, gender, senior citizen (i.e., whether the customer is a senior citizen or not), partner (i.e., whether the customer has a partner or not), and dependents (whether the customer has dependents or not).

Based on the Gini coefficients report and summarization effects of all the features, only senior citizens and gender show a high degree of effects on model prediction outcome. This implies that we can still keep partners and dependents in our model without worrying about model bias.

Regarding gender, the predictor's effect has a low feature value and low mean absolute SHAP values. Hence, by dropping the gender from the model, not only can we prevent any

potential model bias, but also the adverse effects on the model performance by dropping the predictor is lowest. If we drop the predictor, we may lose significant predicting power.

Additionally, age or date of birth predictors is widely used in machine learning models such as credit card approval and insurance underwriting. Hence, we would recommend keeping the senior citizen predictor.

5.2 Further Research

Our team would recommend the marketing analysis at the TelCo continue to investigate the model interpretability report, to identify churn correlations or causes, and to address the issues accordingly. For example, our team has found that if a consumer is with a two-year phone contract, then that consumer is less likely to churn. Hence, the TelCo marketing team could offer better new two-year deals to attract new customers to sign-up for a new two-year contract. However, some factors that contribute to the prediction outcome are not immediately apparent. For example, our team has discovered that if a consumer signed up for paperless billing, it is more likely to churn. The TelCo team cannot derive action such as stop all paperless billing based on the report alone. Instead the TelCo team should deep dive into the causal inference report and discover the unobserved factors, for example, what types of people are more likely to sign-up for paperless billing.