

CHURNS DATA ANALYSIS MODEL



Submitted by:
UJJWAL PRATIK

Contents

ACKNOWLEDGMENT	3
INTRODUCTION	4
ABOUT CHURNS ANALYSIS.....	4
PROBLEM STATEMENT.....	4
DATA ANALYSIS	5
COLUMNS IN THE DATAFRAME.....	5
sample.....	5
OBJECTIVE	5
DATA DESCRIPTION	5
METHODOLOGY	6
METRIC USAGE	6
SYSTEM REQUIREMENTS	7
APPROACH	8
IDENTIFICATION OF POSSIBLE PROBLEM-SOLVING APPROACHES.....	8
TESTING OF IDENTIFIED APPROACH(Algorithms)	9
KEY FOR SUCCESS IN SOLVING PROBLEM UNDER CONSIDERATION	10
CONCLUSION	11
LEARNING OUTCOMES OF THE STUDY IN RESPECT OF DATA SCIENCE	11
REFERENCES	12

ACKNOWLEDGMENT

I sincerely thanks to the Data Trained Faculty for the guidance. They have covered the topics like Machine Language, Python & SQL. Under their guidance I learned a lot about this project. their suggestions and directions have helped in the completion of this project. I had also taken help from YouTube & online videos.

INTRODUCTION

ABOUT CHURNS ANALYSIS

Churn analysis is the process of using data to understand why your customers have stopped using your product or service. Analysing your churn doesn't only mean knowing what your churn rate is. It is about figuring out why customers are churning at the rate they are, and how to fix the problem.

PROBLEM STATEMENT

Customer churn is when a company's customers stop doing business with that company. Businesses are very keen on measuring churn because keeping an existing customer is far less expensive than acquiring a new customer. New business involves working leads through a sales funnel, using marketing and sales budgets to gain additional customers. Existing customers will often have a higher volume of service consumption and can generate additional customer referrals. Customer retention can be achieved with good customer service and products. But the most effective way for a company to prevent attrition of customers is to truly know them. The vast volumes of data collected about customers can be used to build churn prediction models. Knowing who is most likely to defect means that a company can prioritise focused marketing efforts on that subset of their customer base. Preventing customer churn is critically important to the telecommunications sector, as the barriers to entry for switching services are so low.

DATA ANALYSIS

COLUMNS IN THE DATAFRAME

- Customer ID
- gender
- Senior Citizen
- Partner
- Dependents
- tenure
- Phone Service
- Multiple Lines
- Internet Service
- Online Security
- Device Protection
- Tec Support
- Streaming TV
- Streaming Movies
- Contract
- Paperless Billing
- Payment Method
- Monthly Charges
- Total Charges
- Total Charges (Target)

sample

customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	...	DeviceProtection	TechSupport
7590-VHVEG	Female	0	Yes	No	1	No	No phone service	DSL	No	...	No	No
5575-GNVDE	Male	0	No	No	34	Yes	No	DSL	Yes	...	Yes	No
3668-QPYBK	Male	0	No	No	2	Yes	No	DSL	Yes	...	No	No
7795-CFOCW	Male	0	No	No	45	No	No phone service	DSL	Yes	...	Yes	Yes
9237-HQITU	Female	0	No	No	2	Yes	No	Fiber optic	No	...	No	No

OBJECTIVE

- Our main objective is to examine customer data from IBM Sample Data Sets with the aim of building and comparing several customer churn prediction models.
- All the parameters will be analysed through Machine Learning algorithms like Logistic Regression, AdaBoost Classifier, Random Forest Classifier, Decision Tree Classifier, Support Vector Classifier etc which will help to predict.

DATA DESCRIPTION

- The source of data is taken from GitHub.
- Data(https://github.com/dsrscientist/DSData/blob/master/Telecom_customer_churn.csv)

METHODOLOGY

- It gives insights of the dependency of target variables on independent variables using machine learning techniques to determine the churns because it gives the best outcome.
- The independent variable is customer ID, gender, senior citizen, partner, Dependents, tenure, phone service, Multiple Lines, Internet Service, online security, online backup, device protection, tech support, streaming tv, streaming movies, contract, paperless billing, payment method, monthly charges, total charges and dependent Variable is churn.

METRIC USAGE

- a. Logistic Regression.
- b. AdaBoost Classifier.
- c. Random Forest Classifier.
- d. Gaussian NB
- e. Decision Tree Classifier.
- f. Support Vector Classifier

SYSTEM REQUIREMENTS

HARDWARE and SOFTWARE REQUIREMENTS and TOOLS USED

- a) Hardware Requirement:
 - i. Intel core i5
 - ii. 8 GB Ram
- b) Software Requirement:
 - i. Python 3.x with packages:
 - 1. Pandas: Data analysis and manipulation tool
 - 2. NumPy: Provide support for mathematical functions, random number etc.
 - 3. Matplotlib: is a low-level graph plotting library in python that serves as a visualization.
 - 4. Seaborn: is a library mostly used for statistical plotting in python.
 - 5. Scikit-Learn: is an open-source Python library that has powerful tools for data analysis and data mining.

APPROACH

IDENTIFICATION OF POSSIBLE PROBLEM-SOLVING APPROACHES

- **Logistic Regression:** Logistic regression is fast and relatively uncomplicated, and it is convenient for you to interpret the results.
- **AdaBoost Classifier:** s a meta-estimator that begins by fitting a classifier on the original dataset and then fits additional copies of the classifier on the same dataset but where the weights of incorrectly classified instances are adjusted such that subsequent classifiers focus more on difficult cases.
- **Random Forest Classifier:** The Random Forest classifier creates a set of decision trees from a randomly selected subset of the training set. It is basically a set of decision trees (DT) from a randomly selected subset of the training set and then It collects the votes from different decision trees to decide the final prediction.
- **Gaussian NB:** Gaussian NB is based on the Naive Bayes theorem with the assumption of conditional independence between every pair of features given the label of the target class.
- **Decision Tree Classifier:** A decision tree is a flowchart-like tree structure in which the internal node represents feature (or attribute), the branch represents a decision rule, and each leaf node represents the outcome. A Decision Tree consists of Nodes: Test for the value of a certain attribute.
- **Support Vector Classifier:** Support Vector Machine is a discriminative classifier that is formally designed by a separative hyperplane. It is a representation of examples as points in space that are mapped so that the points of different categories are separated by a gap as wide as possible. In addition to this, an SVM can also perform non-linear classification.
- **Cross-Validation-Score:** a model that would just repeat the labels of the samples that it has just seen would have a perfect score but would fail to predict anything useful on yet-unseen data.
- **Grid Search CV:** This function helps to loop through predefined hyperparameters and fit your estimator (model) on your training set.
- **ZScore:** Z-score is also known as standard score gives us an idea of how far a data point is from the mean.
- **Label Encoder:** Label Encoding refers to converting the labels into numeric form.
- **IMB Learn Anaconda:** imbalanced-learn is a python package offering a number of re-sampling techniques commonly used in datasets showing strong between-class imbalance. It is compatible with scikit-learn and is part of scikit-learn-contra projects.
- **Standard Scaling:** Standard Scaler helps to get standardized distribution, with a zero mean and standard deviation of one (unit variance). It standardizes features by subtracting the mean value from the feature and then dividing the result by feature standard deviation.
- **AUC-ROC CURVE:** The AUC-ROC curve tells us visualize how well our machine learning classifier is carrying out. It is one of the most significant evaluation metrics for examining any classification model's performance. It is also called as AUROC (Area Under the Receiver Operating Characteristics).

TESTING OF IDENTIFIED APPROACH(Algorithms)

- a. Train Test Split
- b. Label Encoding
- c. STANDARD SCLAER (SMOTE)
- d. IMB Learn
- e. Logistic Regression
- f. AdaBoost Classifier
- g. Gaussian NB
- h. Decision Tree classifier
- i. Support Vector Classifier
- j. Grid Search CV
- k. Cross Validation
- l. AUC ROC CURVE

KEY FOR SUCCESS IN SOLVING PROBLEM UNDER CONSIDERATION

- Analysed data for any unique values.
- Analysed data for distribution.
- Compared between two columns.
- Checked and removed outliers through zscore method.
- Removed skewness present in the dataset.
- Done Oversampling.
- Cross validate the accuracy score from overfitting.
- Done AUC ROC score for better understanding.
- Hyper Parameter tuning using Grid Search CV.

CONCLUSION

As our conclusion we proclaim that, after checking accuracy score, cross validation, Ensemble Techniques, and checking AUC score we declare Ada Boost Classifier is predicting approx. 80% accuracy is best suited model for our purpose of predicting several customer churn

LEARNING OUTCOMES OF THE STUDY IN RESPECT OF DATA SCIENCE

This study gives me opportunity for lots of learning starting from various types of plotting like histograms, boxplot, scatterplot, line chart and many more graphs. These graphs helped me to analyse different aspects of data like outlier, skewness, correlation etc. It also helped me to learn how to apply various model techniques on data and enable predications.

REFERENCES

- Data trained course videos.
- Google Search.
- YouTube.
- GitHub.
- UCI Machine learning repository.