**MEF UNIVERSITY**

**CUSTOMER SEGMENTATION & CHURN PREDICTION FOR CUSTOMER RELATIONSHIP MANAGEMENT**

**Capstone Project**

**Buğra Balantekin**

**İSTANBUL, 2021**

**MEF UNIVERSITY**

**CUSTOMER SEGMENTATION & CHURN PREDICTION FOR CUSTOMER RELATIONSHIP MANAGEMENT**

**Capstone Project**

**Buğra Balantekin**

**Advisor: Asst. Prof. Dr. Evren Güney**

**İSTANBUL, 2021**

**MEF UNIVERSITY**

Name of the project: Customer Segmentation & Churn Prediction For Customer Relationship Management

Name/Last Name of the Student: Buğra Balantekin

Date of Thesis Defense: dd/mm/yyyy

I hereby state that the graduation project prepared by Buğra Balantekin has been completed under my supervision. I accept this work as a “Graduation Project”.

dd/mm/yyyy

(Asst. Prof. Evren Güney)

I hereby state that I have examined this graduation project by Buğra Balantekin which is accepted by his supervisor. This work is acceptable as a graduation project and the student is eligible to take the graduation project examination.

dd/mm/yyyy

Director

of

Big Data Analytics Program

We hereby state that we have held the graduation examination of Buğra Balantekin and agree that the student has satisfied all requirements.

**THE EXAMINATION COMMITTEE**

Committee Member Signature

1. Your Advisor’s Name
2. Your Advisor’s Name ………………………..

# ACADEMIC HONESTY PLEDGE

I promise not to collaborate with anyone, not to seek or accept any outside help, and not to give any help to others.

I understand that all resources in print or on the web must be explicitly cited.

In keeping with MEF University’s ideals, I pledge that this work is my own and that I have neither given nor received inappropriate assistance in preparing it.



Name Date Signature

Buğra Balantekin 12.07.2021

# EXECUTIVE SUMMARY

CUSTOMER SEGMENTATION & CHURN PREDICTION FOR CUSTOMER RELATIONSHIP MANAGEMENT

Buğra Balantekin

Advisor: Asst. Prof. Evren Güney

JULY, 2021, Number pages (e.g. 45 pages)

Your abstract goes here. All paragraphs in this section follow the format “Normal”. Make sure that the format selected is “Normal”. You can check it from the menu Format>Styles and Formatting that will be opened on the right-hand side of this document.

**Key Words**: Customer Segmentation, Churn Prediction,

# ÖZET

MÜŞTERİ İLİŞKİLERİ YÖNETİMİ İÇİN MÜŞTERİ SEGMENTASYONU VE CHURN TAHMİNİ

Buğra Balantekin

Tez Danışmanı: Dr. Öğr. Üyesi Evren Güney

TEMMUZ, 2021, sayfa sayısı (ör. 45 sayfa)

Projenin Türkçe Özeti. All paragraphs in this section follow the format “Normal”. Make sure that the format selected is “Normal”. You can check it from the menu Format>Styles and Formatting that will be opened on the right-hand side of this document.

**Anahtar Kelimeler**: Müşteri Segmentasyonu, Churn Tahmini

# TABLE OF CONTENTS

[ACADEMIC HONESTY PLEDGE 6](#_Toc76997501)

[EXECUTIVE SUMMARY 7](#_Toc76997502)

[ÖZET 8](#_Toc76997503)

[TABLE OF CONTENTS 9](#_Toc76997504)

[1. INTRODUCTION 1](#_Toc76997505)

[1.1. Customer Segmentation Literature Survey 1](#_Toc76997506)

[1.2. Churn Prediction Literature Survey 1](#_Toc76997507)

[2. DATA 2](#_Toc76997508)

[2.1. Features 2](#_Toc76997509)

[2.2. Exploratory Data Analysis 2](#_Toc76997510)

[3. PROJECT DEFINITION 3](#_Toc76997511)

[3.1. How Do You Apply a Style? 3](#_Toc76997512)

[3.2. Why Use This Style? 3](#_Toc76997513)

[4. METHODOLOGY 4](#_Toc76997514)

[5.1. Citing and Referencing 4](#_Toc76997515)

[5.2. Quotations 4](#_Toc76997516)

[5. RESULTS 6](#_Toc76997517)

[6. BUSINESS APPLICATIONS 7](#_Toc76997518)

[APPENDIX A 8](#_Toc76997519)

[APPENDIX B 9](#_Toc76997520)

[REFERENCES 10](#_Toc76997521)

# 

.

# 1. INTRODUCTION

In some sectors of economy, due to harsh competition, profit margins are narrowed. To expand the profits, businesses need to understand not only their rival’s pricing attitude but also their own marketing campaigns expenditures and these expenditures effects on profits. One way to achieve this is by targeting correct products with correct customers with correct prices. To target customers, customer segmentation should be considered to understand group dynamics. Also, businesses need to understand which customer have tendency to churn so that they can make a final touch to keep these customers onboard.

This project aims to understand the fundamental aspects of Marketing Analytics, Customer Segmentation and Churn Prediction.

## 1.1. Customer Segmentation Literature Survey

Tripathi et al. (2010) has examined the role of customer segmentation in Customer Relationship Management. Authors has compared the main clustering algorithms which are K-means and Hierarchical Clustering. Their findings are not specifically performed on a dataset, but both algorithms are thoroughly inspected for their major drawbacks and advantages.

## 1.2. Churn Prediction Literature Survey

Nie et al. (2011) has gathered data for Chinese banks from a data warehouse, which included 60 million observations and 135 features. Authors have randomly sampled observations and limit their features according to their correlations. They’ve also compared the performance of logistic regression and decision trees, which is also included in this project. According to their measurement logistic regression has yielded better results for churn prediction.

Huang et al. (2012) has also gathered data from a data warehouse in Ireland about telecom customers churn prediction. Their data has more than 800K observations with more than 700 features for each observation, which is a more likely scenario that we can observe in real life, especially in finance sector in terms of feature count. Their project has compared seven different algorithm including linear and non-linear ones.

Keramati et al. (2016) has developed a model for churn prediction with CRISP-DM methodology using Decision Tree algorithms.

Kumar (2008) compared six different algorithms for churn prediction in the research and also used SMOTE (Synthetic Minority Oversampling Technique) for overcoming the imbalance problem in dataset.

# 2. DATA

Dataset for this study is obtained from Kaggle Datasets. [[1]](#footnote-1) It has mainly organized for predicting customer churn in Telecom Industry.

## 2.1. Features

The project dataset has 100 features and 100.000 observations. It has both continuous and categorical variables. It has also lots of missing values especially for categorical variables.

Features of the dataset can be categorized into two segments;

1. Customer’s Demographic Features
2. Customer’s Behavioural Features

List of all variables are not mentioned in this part due to its size. Instead, list of all variables is explained in Appendix A.

## 2.2. Exploratory Data Analysis

Churners and Non-Churners are almost equally distributed in the dataset. (50438 nunchurn - 49562 churn)

# 3. PROJECT DEFINITION

# 4. METHODOLOGY

Since churn prediction is a binary classification problem, multiple classification algorithms are compared. Here’s a list of compared algorithms;

* Logistic Regression
* Decision Tree Classifier
* Random Forest Classifier
* Gaussian NB Classifier
* KNN classifier
* XGBoost Classifier
* Adaboost Classifier
* Voting Classifier
* Light GBM
* MLP Classifier

After dropping demographic features from dataset, as they’re heavily categorical and has less explain ability for churn prediction from business perspective, FancyImputer package of Python is used to fill these values. Main method to fill missing values is MICE (Multiple Imputation by Chained Equation). KNN imputation method could not be used because of the Google Colab or Local Python IDE (VS Code) kept crashing due to RAM size.

To select the best variables for predicting churn, couple of methods are used;

* Low variance features are dropped
* Features that are highly correlated (greater than 95 pct) are dropped to avoid multicollinearity
* Automated feature selection with Boruta Package

Even though the dataset is not constructed for customer segmentation, Churn column will be dropped to segment the customers using the given features. Effective Customer Relationship Management for any company is based on differentiation of customers regarding their behavior. For customer segmentation, K-means clustering will be used with elbow method and silhouette analysis for selecting the optimal K number. After the execution of algorithm, cluster analysis will be made with visualizations. Also, two-step clustering algorithm will be used for cluster analysis. This part will be made with IBM SPSS Modeler software.

# 5. RESULTS

Time and metric comparison of algorithms will be here…

# 6. BUSINESS APPLICATIONS

# APPENDIX A

* 1 rev\_Mean Mean monthly revenue (charge amount)
* 2 mou\_Mean Mean number of monthly minutes of use
* 3 totmrc\_Mean Mean total monthly recurring charge
* 4 da\_Mean Mean number of directory assisted calls
* 5 ovrmou\_Mean Mean overage minutes of use
* 6 ovrrev\_Mean Mean overage revenue
* 7 vceovr\_Mean Mean revenue of voice overage
* 8 datovr\_Mean Mean revenue of data overage
* 9 roam\_Mean Mean number of roaming calls
* 10 change\_mou Percentage change in monthly minutes of use vs previous three month average
* 11 change\_rev Percentage change in monthly revenue vs previous three month average
* 12 drop\_vce\_Mean Mean number of dropped (failed) voice calls
* 13 drop\_dat\_Mean Mean number of dropped (failed) data calls
* 14 blck\_vce\_Mean Mean number of blocked (failed) voice calls
* 15 blck\_dat\_Mean Mean number of blocked (failed) data calls
* 16 unan\_vce\_Mean Mean number of unanswered voice calls
* 17 unan\_dat\_Mean Mean number of unanswered data calls
* 18 plcd\_vce\_Mean Mean number of attempted voice calls placed
* 19 plcd\_dat\_Mean Mean number of attempted data calls placed
* 20 recv\_vce\_Mean Mean number of received voice calls
* 21 recv\_sms\_Mean Mean number of received sms messages
* 22 comp\_vce\_Mean Mean number of completed voice calls
* 23 comp\_dat\_Mean Mean number of completed data calls
* 24 custcare\_Mean Mean number of customer care calls
* 25 ccrndmou\_Mean Mean rounded minutes of use of customer care calls
* 26 cc\_mou\_Mean Mean unrounded minutes of use of customer care (see CUSTCARE\_MEAN) calls
* 27 inonemin\_Mean Mean number of inbound calls less than one minute
* 28 threeway\_Mean Mean number of three way calls
* 29 mou\_cvce\_Mean Mean unrounded minutes of use of completed voice calls
* 30 mou\_cdat\_Mean Mean unrounded minutes of use of completed data calls
* 31 mou\_rvce\_Mean Mean unrounded minutes of use of received voice calls
* 32 owylis\_vce\_Mean Mean number of outbound wireless to wireless voice calls
* 33 mouowylisv\_Mean Mean unrounded minutes of use of outbound wireless to wireless voice calls
* 34 iwylis\_vce\_Mean Mean number of inbound wireless to wireless voice calls
* 35 mouiwylisv\_Mean Mean unrounded minutes of use of inbound wireless to wireless voice calls
* 36 peak\_vce\_Mean Mean number of inbound and outbound peak voice calls
* 37 peak\_dat\_Mean Mean number of peak data calls
* 38 mou\_peav\_Mean Mean unrounded minutes of use of peak voice calls
* 39 mou\_pead\_Mean Mean unrounded minutes of use of peak data calls
* 40 opk\_vce\_Mean Mean number of off-peak voice calls
* 41 opk\_dat\_Mean Mean number of off-peak data calls
* 42 mou\_opkv\_Mean Mean unrounded minutes of use of off-peak voice calls
* 43 mou\_opkd\_Mean Mean unrounded minutes of use of off-peak data calls
* 44 drop\_blk\_Mean Mean number of dropped or blocked calls
* 45 attempt\_Mean Mean number of attempted calls
* 46 complete\_Mean Mean number of completed calls
* 47 callfwdv\_Mean Mean number of call forwarding calls
* 48 callwait\_Mean Mean number of call waiting calls
* 49 churn Instance of churn between 31-60 days after observation date
* 50 months Total number of months in service
* 51 uniqsubs Number of unique subscribers in the household
* 52 actvsubs Number of active subscribers in household
* 53 new\_cell New cell phone user
* 54 crclscod Credit class code
* 55 asl\_flag Account spending limit
* 56 totcalls Total number of calls over the life of the customer
* 57 totmou Total minutes of use over the life of the cus
* 58 totrev Total revenue
* 59 adjrev Billing adjusted total revenue over the life of the customer
* 60 adjmou Billing adjusted total minutes of use over the life of the customer
* 61 adjqty Billing adjusted total number of calls over the life of the customer
* 62 avgrev Average monthly revenue over the life of the customer
* 63 avgmou Average monthly minutes of use over the life of the customer
* 64 avgqty Average monthly number of calls over the life of the customer
* 65 avg3mou Average monthly minutes of use over the previous three months
* 66 avg3qty Average monthly number of calls over the previous three months
* 67 avg3rev Average monthly revenue over the previous three months
* 68 avg6mou Average monthly minutes of use over the previous six months
* 69 avg6qty Average monthly number of calls over the previous six months
* 70 avg6rev Average monthly revenue over the previous six months
* 71 prizm\_social\_one Social group letter only
* 72 area Geogrpahic area
* 73 dualband Dualband
* 74 refurb\_new Handset: refurbished or new
* 75 hnd\_price Current handset price
* 76 phones Number of handsets issued
* 77 models Number of models issued
* 78 hnd\_webcap Handset web capability
* 79 truck Truck indicator
* 80 rv RV indicator
* 81 ownrent Home owner/renter status
* 82 lor Length of residence
* 83 dwlltype Dwelling Unit type
* 84 marital Marital Status
* 85 adults Number of adults in household
* 86 infobase InfoBase match
* 87 income Estimated income
* 88 numbcars Known number of vehicles
* 89 HHstatin Premier household status indicator
* 90 dwllsize Dwelling size
* 91 forgntvl Foreign travel dummy variable
* 92 ethnic Ethnicity roll-up code
* 93 kid0\_2 Child 0 - 2 years of age in household
* 94 kid3\_5 Child 3 - 5 years of age in household
* 95 kid6\_10 Child 6 - 10 years of age in household
* 96 kid11\_15 Child 11 - 15 years of age in household
* 97 kid16\_17 Child 16 - 17 years of age in household
* 98 creditcd Credit card indicator
* 99 eqpdays Number of days (age) of current equipment
* 100 Customer\_ID

# APPENDIX B

# REFERENCES

Use APA styling for citations

<http://www.bibme.org/apa>

1. <https://www.kaggle.com/abhinav89/telecom-customer> [↑](#footnote-ref-1)