EASTERN UNIVERSITY

Machine Learning Project Proposal

DTSC691: Applied Data Science

John Behler

Project Overview

For my capstone project, I will be looking at customer banking data to analyze customer credit card churn and see which type of customers stay with the bank for a longer duration. I will use machine learning to create logistical and linear models to identify which factors have the most impact on customer retention. I will also use tableau for visualizations to show trends from my analysis. I want to see what variables are statistically significant in predicting how long a customer stays at a bank. This is important for a bank to know so they can adjust their strategy to keep customers at the bank because that can save the bank money.

Project Goals

Purpose

The purpose of this project is to find ways to keep bank customers to stay with the bank for a longer duration of time. Once we look more in depth at the data we will be able to see trends and patterns for which demographics tend to have a longer tenure. When we can identify the causes of what causes churn, the business owner can find ways to keep customers with the bank for a longer period.

Project Focus

I will be using data analysis, predictive modeling, and creating visualizations to predict customers credit card tenure. There are many different variables such as country, credit card score, age, and gender which may play a role in how long a person decides to keep their credit card with the bank. It is important for a bank to know what factors lead customers to stay longer so they can implement strategies.

Specific Goals

I will use data visualization in tableau to create graphs to show to business owners. Also I will use regression models to see which factors are important predictors in determining how long customers keep their card with the bank.

Expected Outcomes

I will hopefully create a model that will be able to accurately predict how long a customer will hold onto their credit card at the bank based on their demographics. Being able to look at which variables are important predictors will help give insight to the bank so they can change their marketing strategies. Hopefully we will be able to learn some statistically significant information that will help us determine what causes someone to stay with the bank longer.

Project Description

Project Objective and Scope

The scope of this data is just this one bank because we only have data from here. So we won't be able to confidently say that the information we get from our analysis will apply to all banks. We could maybe implement our model for other banks but each bank might have a different conclusion. The objective of this project is to dive deeper into the variables that affect bank retention. This is important information for a bank to be aware of so they can adjust their marketing strategies.

Data Description

I have found my data source on kaggle which has 18 columns and 10,000 rows. There is categorical and numerical data. We will use tenure as the response variable and the other variables as the predictors.The response variable is tenure, as this is what we are trying to predict.

Exploratory Data Analysis

In Jupyter I will look at the distribution of the variables when selecting which features to include in my model. I will use Tableau to create visualizations to show the relationship between different variables in the bank data and customer churn. This will be helpful in order to show a relationship to be able to show a manager. Graphs and visualizations are very important because it is easier to show and understand the data for someone who might not have data science experience.

Data Preparation and Cleaning

I will go through the data to make sure to deal with missing data. When this happens I will have to decide to remove the entire row or to input missing data with the mean or median if the data is numerical. I will also want to look for outliers in the data and remove them if I feel it will not benefit my model. I will also check to see that categorical data is spelled correctly so each category is consistent for grouping. Once I look at the data more clearly, I might want to feature engineer a variable if I think it could be important. I may have to convert categorical data to numerical in order to use it in a linear or logistic regression.

Model Training

I will look at different sorts of models to predict customer churn. I will run logistic models for categorical variables and linear regression models for numerical variables. I will choose variables that appear to have a relationship with bank tenure. I can do this by looking at the correlation between the variables. I will also want to look at the distribution of variables.

Model Evaluation

I will use key statistics to determine the model's efficiency. When identifying if a model is efficient and if variables are statistically significant, I will look at the RMSE, r^2, p-value, f1 score, precision, recall, etc to determine how good the model is. If a model is inefficient I will create a new model with adjustments and compare the metrics.

User- Interface Integration

I will create a tableau Dashboard for interaction. I will incorporate different variables and graphs where a user can interact with customer bank churn. Visualizations are a good way to explain your findings for someone who doesn't have the data science background and can't fully understand the models.

Capstone Complexity

This project will meet the capstone level project requirements because I will be using techniques from many different courses that I have learned. First off, I will be cleaning the data to look for missing data. Then I will be creating machine learning techniques to look at statistically significant variables. This will help us predict how long a customer will stay at the bank. Then I will use tableau to create visualizations to be able to present my findings. I will use many techniques learned from different classes to help me with this capstone project.

Software

I will be using Jupyter (Python) to work with data cleaning, data analysis, and machine learning. Then I will use tableau to create visualizations and create a tableau dashboard for interaction. Once I finish the machine learning and visualizations, I will use powerpoint to create my presentation. Then I will record myself presenting using my macbook air.

Project Completion Plan

- ### Week 1

  I read the syllabus, course handbook, set up google drive, took the quiz, and read the documents that would relate to the capstone project. I decided I wanted to do a machine learning project and then selected Bank Customer Churn data for my final project. Then I submitted my completed project topic form.

- ### Week 2

  During week two, I worked on my project proposal and connected with my mentor. I looked at the project proposal guideline and rubric and worked on filling out my template. I looked over the data to see what I would need to do in order to complete this project which helped me fill out this project proposal.

- ### Week 3

  For week three, I will import the data into Jupyter (Python) and will do exploratory data analysis and look at each variable and gain more information about them. Then I will data clean which will include removing duplicates, looking at missing data, making sure categorical data is spelled correctly, etc. Then I will look at correlations between the variables and I will decide which I will decide to incorporate in my model. I will create logistic and linear regression models. I will look at the metrics to evaluate the efficiency of the model and make adjustments if necessary.

- ### Week 4

  During week four, I will import the data into Tableau and create data visualizations. I will look at different variables and make different graphs that I will easily be able to convey my analysis to a manager.

- ### Week 5

  During week 5, I will put the project together. I plan on making a powerpoint with my analysis. This will include my machine learning models and tableau visualizations. I will incorporate graphs and visuals in my presentation.

- ### Week 6

  During week 6, I will record my project. I will use the powerpoint I created to present my analysis. I will use my macbook air to record my presentation.

- ### Week 7

  I hope to have my capstone project done by the end of week 6 but I am giving myself an extra week for flexibility in case I need to spend more time on a different topic.

Presentation Plan

I will create a powerpoint and include slides with my findings from my code and visualizations. I will use my macbook air to record my presentation.

Resources

I will use brightspace as a resource to go back to videos and find documentation for when I need to revisit a topic to complete my project. If I can’t find something on brightspace I will use google and youtube. I understand this project is supposed to mainly be independent but I will also use my mentor if I need help with something that I can not figure out.