1. Introduction:

AndPlus+ is a business consultancy focused on digital transformation and provides various solutions focused on software design and development which increases experience, decreases cost and create unique platforms from where users can extract value. Based in Southborough, MA, AndPlus+ provides where solutions ranging from Web applications, mobile and IoT solutions for their clients which are recorded in tickets representing the various details about the solution that is being provided. The company wants their ticketed project data for various clients analyzed to resolve to problems that need to resolve, which is forked into two phases. Each second phase is built on the first, details of which are explained in the upcoming section.

1. Problem Description:
   1. Phase 1: In the Phase 1 of the Project the company wants us to analyse their ticketing data. The ticketing data has various, potentially useful data attributes, details of which are elaborated in the next section. A model is the expected outcome of the Phase, which is expected to help the company decide whether a ticket of the project falls within the normal range of hours or not. This will aid the company to decide whether or not they should commit to the project or not. Also, it will help the company allocate resources to the project/ticket accordingly.
   2. Phase 2: A Phase built based on the first one, we are supplied with additional internal data coupled data from various external factors that help determine whether a client will opt for another set of services from the company or not. Simply put, we need to decide why a client is deciding for or against choosing AndPlus services and serve as a starting point to develop strategies to handle the same.

1. Features/Variables Definition:
   1. Phase 1:

For phase 1 of the project, our goal is to be able to estimate if a given ticket will be completed within a normalized range of time. There are about 9000 records that AndPlus would provide us for this problem and it will be available in CSV format. The data that will be used consist of information of time spent on each customer ticket, the complexity of each ticket, number of use cases for each ticket,etc. Detailed description of available features is here:

Independent variables:

1. ID - Unique data instance identifier for each ticket
2. Group - Anonymized text code representing a client/customer
3. Score - Integer value representing ticket complexity based on fibonacci scale
4. User Story - Boolean value indicating presence or absence of a user story
5. Use Cases - Integer value showing number of use cases for each ticket
6. Content Length - Count of text characters in the user story (if present)
7. Time Spent - Number of hours taken to complete the ticket
8. Time to Start - Number of hours passed between ticket arrival till start of ticket solving

Dependent variable:

1. In Normal Range - Boolean value indicating if a ticket is completed within a normalized completion time range

These features are readily available to be used for our analysis and model building. However, based on our research and brainstorming, we thought of a bunch of additional data that if received, could be added to our model and perhaps enhance our model’s prediction capability. This includes Size of the client company in terms of revenue and number of employees, industry/domain of the client, size of team working on each ticket, timeframe/deadline attached to each ticket, priority associated to each ticket and month when ticket is requested. These features are still in discussion with our mentors from AndPlus and we are not sure if these would be incorporated within our project.

b. Phase 2:

Phase 2 of our project is very open-ended, has a wider scope and could require extensive data collecting and pre-processing. For this, there is no data that is available currently. Nonetheless, data from Phase 1 can be incorporated. Some other data features that might be useful are data from past repurchases, information about past projects and if they were completed within expected timeframe and assigned budget. Other data from past projects could be used such as customer satisfaction level, frequency of communication with the customer, type of communication with the customer i.e. email, phone, in-person meetings. This data could also include future needs and budget of the client. Apart from these, there could be numerous external factors such as macroeconomic trends, weather changes, changes in federal laws, political changes and so on.

1. Methodology for problem solving:

We aim to solve this problem by building a data science pipeline (OSEMN) such as:

O: Obtain Data

S: Scrubbing & Cleaning Data

E: Exploring data using statistics and data visualizations

M: Modeling on Data

N: Interpretation

To elaborate on this:

For understanding the data distribution, we will be using visual aids such as build scatter plots, histograms, box-plots and analyze spread of data. Also, using certain statistical tests to also understand the data. There will be outlier detection and elimination, handling of missing data and feature engineering.

For the first phase of the project, it is a bi-classification and hence we will be applying different machine learning models such as Logistic regression, Naive Bayes Classifier, Decision trees, etc.

1. Expected outcome and Deliverables: Build a model that predicts whether the given ticket is in normalized range by the end of phase 1. Analysing the data and come up with a descriptive analysis regarding the declination of returning customers of the company by the end of phase 2.
2. Resources to be used:

i) Languages and Tools: Python, IPython Notebook, PyCharm, R, Rstudio, Google Colab.

ii) Preprocessing: <https://www.researchgate.net/publication/319990923_Review_of_Data_Preprocessing_Techniques_in_Data_Mining>

iii) Elements of Stats Learning: <https://web.stanford.edu/~hastie/ElemStatLearn/>

iv) Datasets: <http://www.pewresearch.org/download-datasets/> and <https://www.census.gov/econ/geo-us.html>

v) ISLR :<https://github.com/JWarmenhoven/ISLR-python>

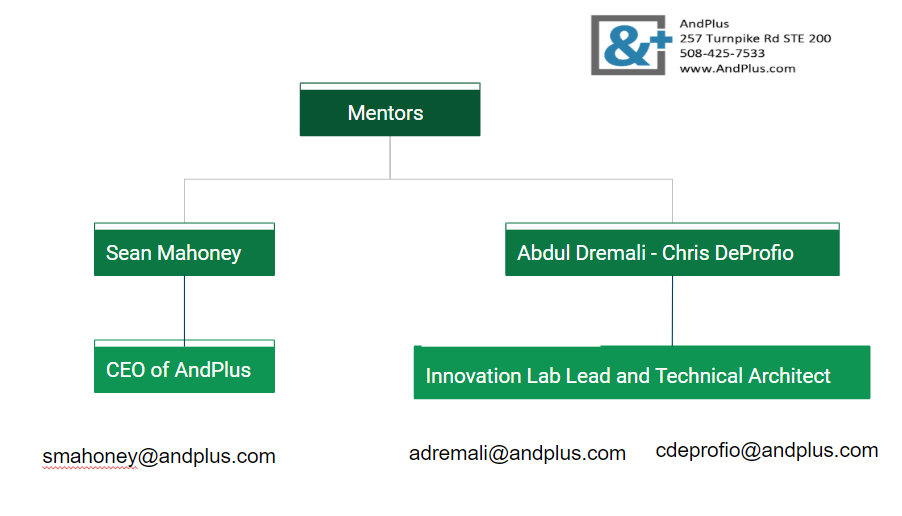
vi) Data Visualisation: <http://www.storytellingwithdata.com/book/>

vii) Data Interpretation: <http://www.stat.cmu.edu/~brian/701/notes/paper-structure.pdf>

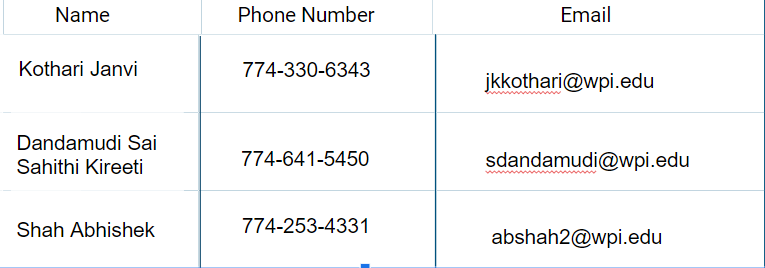
vii) Data Science for Business: <http://shop.oreilly.com/product/0636920028918.do>

ix) Research Papers:

1. Improved customer Churn and Retention Decision Management using Operations Research approach: <https://www.researchgate.net/publication/312131155_Improved_Customer_Churn_and_Retention_Decision_Management_Using_Operations_Research_Approach>
2. AndPlus’s Analytical Approach to Account Management(Internal paper within organisation)
3. Contact Information:
4. Sponsor



1. Group members



Attachments:

Janvi Kothari Resume