

Project Title: CRM Data Cleansing

Sponsoring Organization: Berkadia Commercial Mortgage LLC

Sponsoring Individual with Title: Gary Mumford, Vice President of Data Governance

Sponsoring Individual Phone & Email: 801-285-9643, gary.mumford@berkadia.com

Faculty Advisor with Email: Professor Rohit Aggarwal, rohit.aggarwal@eccles.utah.edu

Submission Date: June 2019

Project Completion Date: June 2019

Project Presentation: December__ April__ July X

Semester Taking IS-6596-003: Summer, 2019

Will All Team Members Take IS-6596-003 Together: Yes X No__

Project Objectives and Requirements

Overview:

- Many companies struggle with CRM data accuracy. This is because good CRM data decays rapidly. Contacts change roles within companies, leave their current company, change phone numbers, or move. This constant change in modern businesses require data to be constantly updated in order to stay valid. Having valid data allows Sales & Marketing teams to be more effective and is in direct correlation to increases in revenue.
- Considering above factors, an approximate percentage change of 2% per month and 25% per year is observed by Berkadia in the CRM data. The project is inclined towards working on refining the CRM accounts and contacts data.
- The Primary problem of the dataset is duplicate entries.
- We will not be directly connected to the CRM database instead we would be given data in CSV or Excel format.

Below is our solution approach:

1. Initial Sampling

- Around 57190 CRM data would be provided giving the contact (197697) and account details in an Excel sheet.
- Build a sample representing the entire data set.

2. Validation

- Testing the sampled data to evaluate the accuracy using different resources (Google, LinkedIn).
- Create dashboards to identify data quality metrics and also devise algorithms for detecting inaccurate and duplicate data.

3. Algorithm selection

- How a specific algorithm improves the data quality metrics.
- Automation of the entire process to clean and provide confidence score for each duplicate or inaccurate data.

Technologies Involved:

Spark(recommended), Python & Machine Learning libraries of Python, Tableau or Power BI for Data visualization.

Deliverables:

- The Dashboard.
- Description of the algorithm applied.
- The actual source-code.

Milestones:

Project Working Hours: 10 Hours/Week

Milestones	Timeline (Expected)
Initial Sampling	End of January 2019
Research to list the algorithms	End of February 2019
Validate the selection Process	Mid-March 2019
Finalize an Algorithm - Validation and Dash-boarding	Mid-April 2019
Implementation	May 2019
Project Presentation	June 2019

Dates:

The submission would be made by June 2019.

Risks and Assumptions

- Meeting set-up & Meeting dynamics will be discussed as per availability and convenience,
- Statistical Sampling should be done with complete ownership.
- Training will not be imparted and learning by students will be driving the project. Students will be taking advise from the Faculty Mentor.

Team Members with Team Role, Email and Student Number

Abhishek Anney: Equal Participation in completing the project deliverables and meeting the deadlines.

Sana Kaur: Equal Participation in completing the project deliverables and meeting the deadlines.

Sukrit Sen: Equal Participation in completing the project deliverables and meeting the deadlines.

Bhuvananjali Challagalla: Equal Participation in completing the project deliverables and meeting the deadlines.

Final Approval Signatures

Gary Mumford:



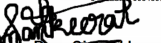
Date: 12/19/2018

Abhishek Anney:



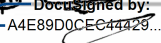
Date: 12/20/2018

Sana Kaur:



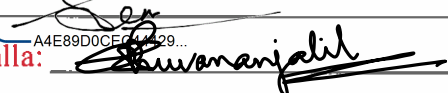
Date: 12/20/2018

Sukrit Sen:



Date: 12/20/2018

Bhuvananjali Challagalla:



Date: 12/20/2018

Faculty Advisor (Rohit Aggarwal):



Date:

TIR (Dave Norwood):



Date: 1/13/2019