Carnegie Mellon University

Project Progress Report

MARCH 3, 2021

Frank Kovacs, Ning Gao, Pragya Jain, Wonil Lee

Agenda

- Introduction
- Background
  - Company Overview
  - Technical Knowledge
- Project Overview
  - Problem
  - Solution
  - Benefit
- Next Steps
- Q&A

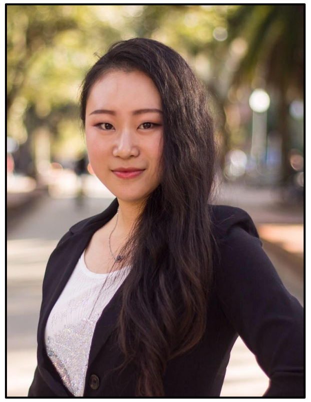
Introduction

Team

Frank Kovacs

- CMU Statistics & Machine Learning ‘19
- Software & Data Research
- Research with Delphi COVIDcast and ISLE

Ning Gao



- Georgia Tech Industrial & Systems Engineering ‘20
- Research with NSF LeapHi Program
- Past work experience in the telecom industry

Pragya Jain

- Past work experience in the insurance industry
- Associate Actuary
- B.E. from NSIT, New Delhi

Wonil Lee

- Past work experience in Consulting (2+ years)
- CMU Tepper & Statistics ‘18
- R, SQL, and Python

Faculty Advisor

Valerie Ventura

- Associate professor in the Department of Statistics and Data Science @ CMU
- Affiliated faculty in the Machine Learning Department, The Center for the Neural Basis of Cognition (CNBC)
- Graduate advisor for the Program in Neural Computation (PNC) at the CNBC
- Ph.D. in Statistics from the University of Oxford

Background

NPD Group Overview

- NPD Group is a **Market research company**
- “Raw data assets into insights”
- Specialize in general merchandise and food service
- Market leader
    - **8B+** B2B transactions / yr

Technical Knowledge

- **Stakeholders**
  - Andrew Dombrowski - Director of Data Science **(SPOC)**
  - Jane Ahlfors - Director of Market Research
  - Tom Poulos - Head of Global Strategy
- **Technical Knowledge**
  - Competent in statistical analysis
  - Exploring anomaly detection

Problem

Objective & Scope

- “...explore using unsupervised learning methods to help identify common data collection errors to help guide further analyst review.“
- **Goals**
  - Identify common data collection errors
  - Facilitate further data analyst review
  - Automate data error flagging processes

Source: Capstone Handout

Main Issue

- **data corruptions**
  - type, price, quantity
- **missing values**
- **unexpected changes in data structure or values**
  - sales data, receipts

Why is this a problem?

- **Inefficiency**
  - Unidentified errors -> damage the efficiency of Data Analysis process
  - Lack of automated error detection in large datasets -> decrease productivities of Data Analyst Team
- **Brand Equity**
  - The core value in the market research industry is the reliability of the data collected
  - Non-error-free deliverables to NPD client -> hurt client satisfaction rate/ loyalty

Solution

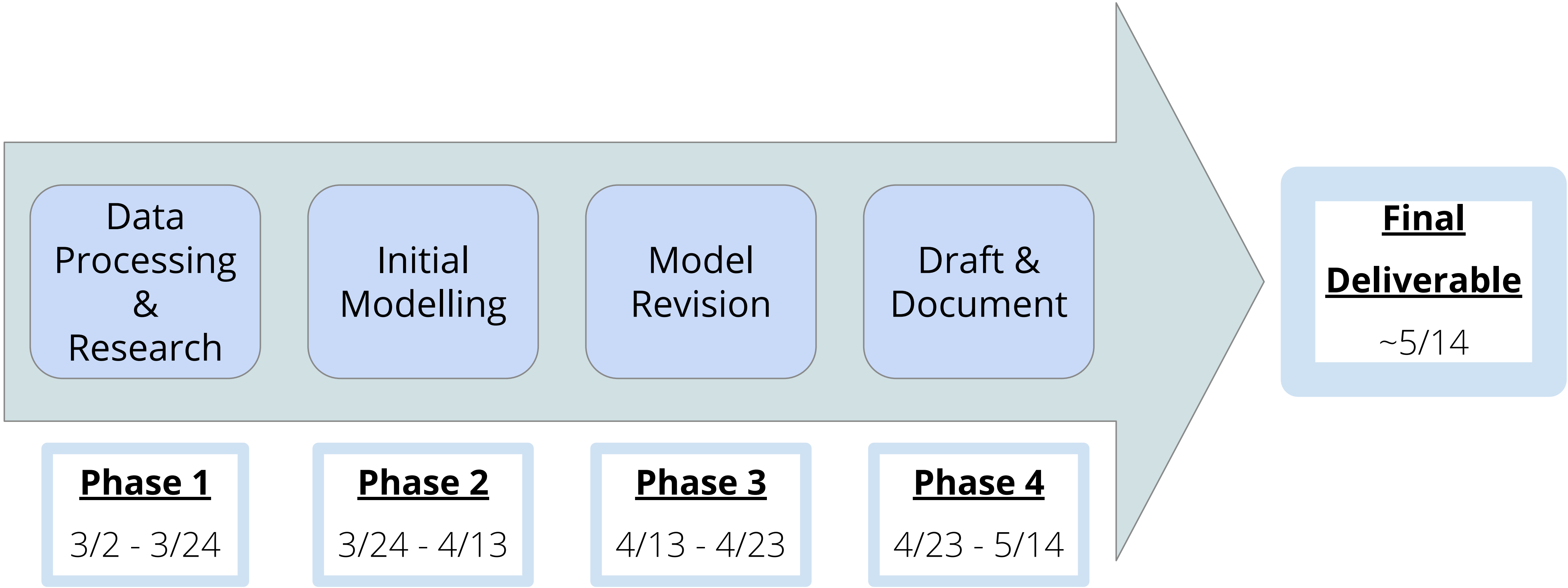
DATA

- **Point of Sale Data**
  - Does not contain consumer information
- **Consumer Surveys**
  - Contains some demographic information
- **Receipt Data**
  - RecieptPal program: Voluntary disclosure with rewards
  - 6-7 years of data

Our role

- Identify issues in time series data
- Prescribe high-level remedies for data issues
- Automate error detection with unsupervised ML method
- Design scalable, easily adjustable algorithm
- Provide recurring weekly/monthly error output table

Project Timeline



Data Processing & Research

Initial Modelling

Model Revision

Draft & Document

Phase 1

3/2 - 3/24

Phase 2

3/24 - 4/13

Phase 3

4/13 - 4/23

Phase 4

4/23 - 5/14

Final Deliverable

~5/14

Benefit

Benefit

- Streamlined error detection process
- Automated reports standardize team-based analysis
- Lessen Redundancy
- Data analysis process made more efficient

Next Steps

Carnegie Mellon University

Next Steps

- Expecting to receive data this week
- Existing error flags and classification labels
- No manual adjustments were made to the data to tackle impact of Covid-19
- Research on suitable anomaly detection methods

Q&A

Contact Information

- Professor Ventura: vventura@andrew.cmu.edu
- Frank : fkovacs@andrew.cmu.edu **(Single Point of Contact)**
- Ning : ningg@andrew.cmu.edu
- Pragya : pragyaj@andrew.cmu.edu
- Wonil : wonillee@andrew.cmu.edu

THANK YOU!