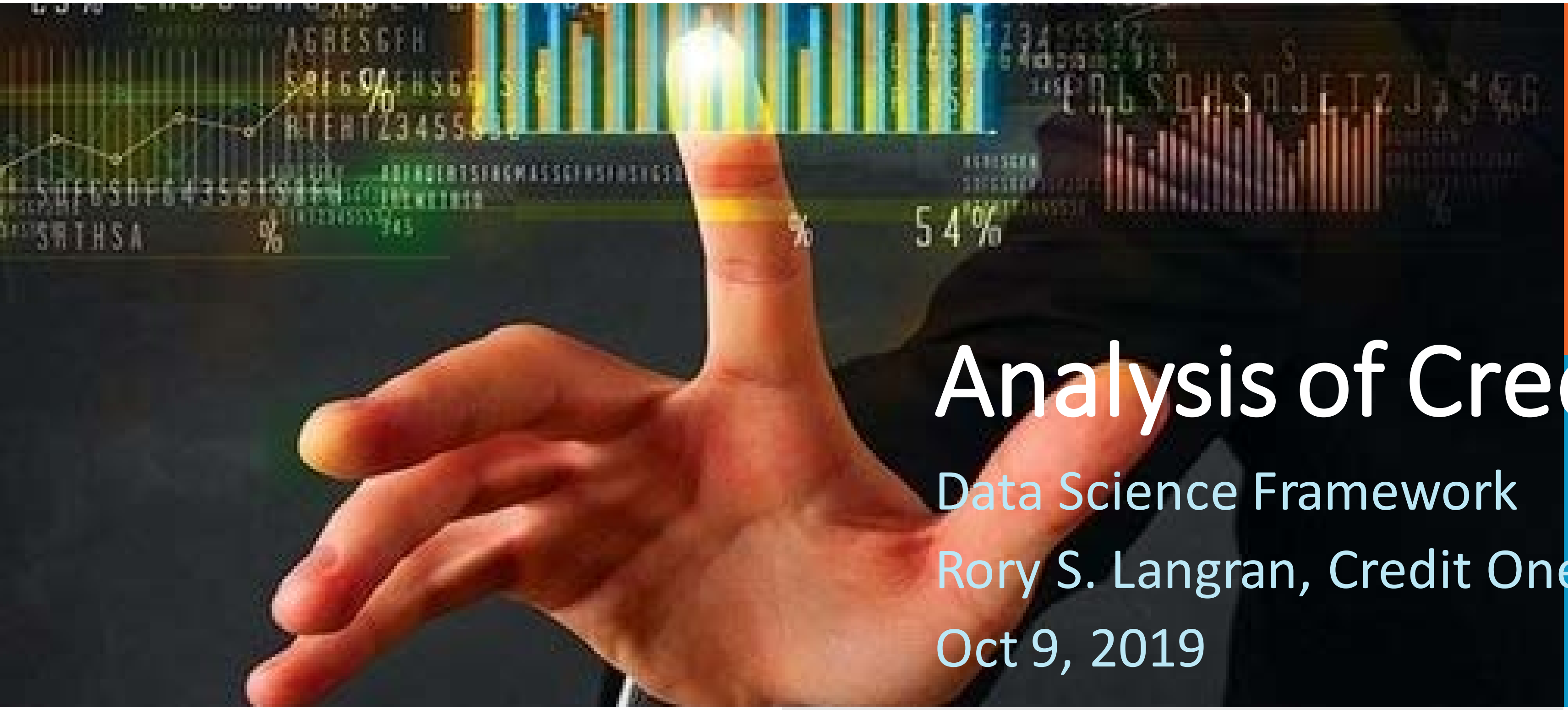
Credit One LLC



Analysis of Credit Default Data

Data Science Framework

Rory S. Langran, Credit One LLC

Oct 9, 2019

Agenda

01 **Background and Goals**

Project Context and Desired Outcomes

02 **Data Description**

An Overview of the Dataset

03 **Data Science Framework**

Our Approach to Managing the Project

04 **Data Management**

Identification of Data Issues

05 **Data Process Flow**

Diagram of How Data Will be Processed

06 **Additional Insights**

Other Ideas Based on Review of Data

01

BACKGROUND AND GOALS

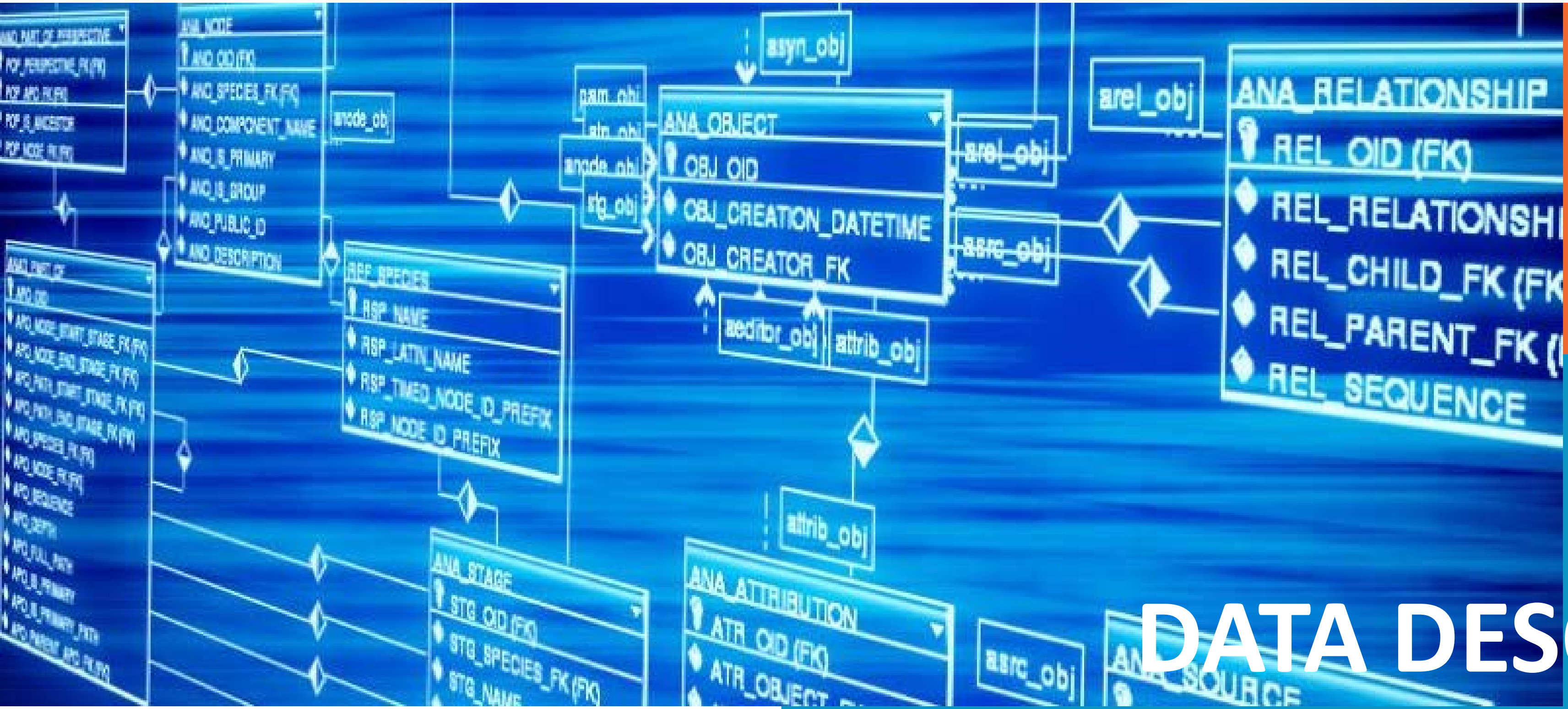
Project Context and Desired Outcomes

Project Background

- Over the past year, there has been an increase in the number of customers who have defaulted on loans from various partners
- Credit One, as their credit scoring service, could risk losing business if the problem is not solved right away
- Management has engaged with the Data Science team in order to design and implement a creative, empirically sound solution for predicting credit default

Project Goals

- Conduct an initial exploration of the current data feed
- Define the business problem using a Data Science framework
- Understand how Data Science will be used to create a model that will more accurately predict credit default
- Design and implement a data process flow model



02

DATA DESCRIPTION

An Overview of the Dataset

Data Description

Background: Customer profile with balance and payment details in *.csv file format

Period: April to September 2005

Records: 30,000

Dependent Variable: Binary loan default indicator for next month – “1” default or “0” no default

Independent Variables: 23 total

Variable Categories:

1) Customer Profile (5): Credit Limit, Gender, Education Level, Marital Status, Age
2) Payment Status (6): Indicator for on-time payment, no payment, or # months late from April to September 2005
3) Balance Amount (6): Balance of account for each month from April to September 2005
4) Payment Amount (6): Payment for each month from April to September 2005

Data Attributes Details

Customer Profile Attributes (5):

- "LIMIT_BAL" – Integer data type; amount of given credit (NT dollars)
- "SEX" – Factor data type; "1" male / "2" female
- "EDUCATION" – Integer data type; "1" graduate school / "2" university / "3" high school / "0, 4, 5, 6" all other school
- "MARRIAGE" – Integer data type; "1" married / "2" single / "3" divorced / "0" all others
- "AGE" – Integer data type; current age of account holder

Payment Status Attributes (6):

- "PAY_0, PAY_2, PAY_3, PAY_4, PAY_5, PAY_6" – Payment indicator for six month period from Apr 2005 through Sep 2005. Integer data type; "-2" no payment required / "-1" paid in full / "0" partial payment / "1,2,3,…" number of months payment is past due

Data Attributes Details

Balance Amount Attributes (6):

- "BILL_AMT1, BILL_AMT2, BILL_AMT3, BILL_AMT4, BILL_AMT5, BILL_AMT6" – Billing statement amount for six month period from Apr 2005 through Sep 2005. Integer data type.

Payment Amount Attributes (6):

- "PAY_AMT1, PAY_AMT2, PAY_AMT3, PAY_AMT4, PAY_AMT5, PAY_AMT6" – Monthly payment amount for six month period from Apr 2005 through Sep 2005. Integer data type.

Default Indicator:

- "default payment next month" – Binary data type classifier; Status of account indicating if it has defaulted "0" or not defaulted "1"

Credit One LLC

03

DATA SCIENCE FRAMEWORK

Our Approach to Managing the Project

Framework One - Zumel and Mount (5 Steps)

1 Define Goals – Understand desired customer outcomes and evaluate feasibility based on the available data

2 Manage Data – Conduct exploratory data analysis and determine areas to improve quality and any other data preparation

3 Develop Model – Design and test predictive models and compare results - refine models for higher accuracy and lower errors

Framework One - Zumel and Mount (5 Steps)

4 Present Results – Present results of the selected model to customer and solicit feedback about confidence in results – may require additional rework based on feedback

5 Deploy Model – Provide knowledge transfer to final model custodian and ensure thorough understanding of structure and process for follow-on maintenance or enhancements

Credit One LLC

04

DATA MANAGEMENT

Identification of Data Issues

Data Management

- Ground truth data will be sourced from *.csv file
- We will train and test predictive models based on ground truth data
- We can leverage *.csv / *.txt / *.xlsx data files or data tables for unclassified data for final testing
- Upon implementation, Python model should be setup to receive data from a database table, such as SQL (PostgreSQL or Microsoft), MySQL (Oracle) , or NoSQL (Apache Cassandra)
- Data ETL and preparation and wrangling will be conducted via Python script(s) and final output to database table (preferred)

Data Management – Any Known Issues

- Verify that final data is available through database
- Test and validate access to database table for read / write
- Ensure that the database table datatypes are the same in *.csv file
- Determine if there are any other available data attributes, that can be leveraged such as credit score, liabilities, etc.
- Evaluate other data sources that can be leveraged in addition to the bank details, such as credit reporting data
- Identify and address any Personally Identifiable Information (PII) issues that could occur with the data collection and reporting

Credit One LLC

05

DATA PROCESS FLOW

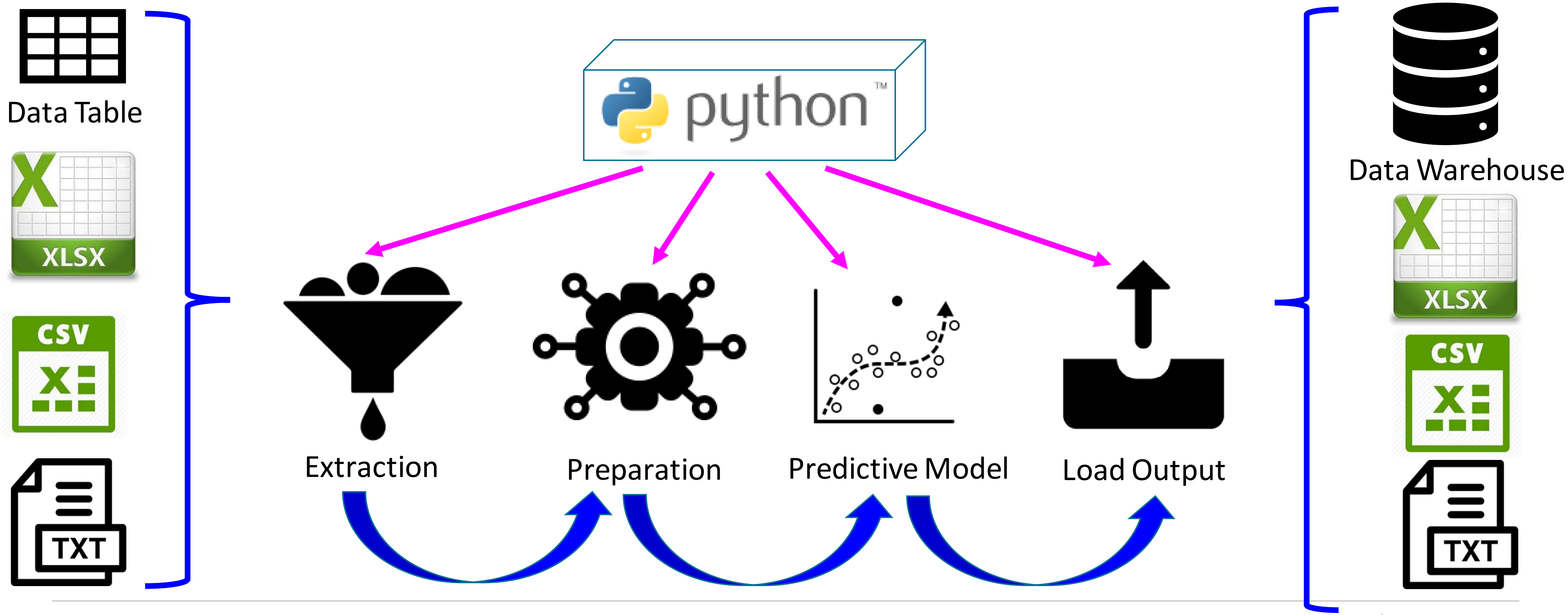
Diagram of How Data Will be Processed

Data Process Flow

Possible Data Sources

Data Extraction, Transformation, and Loading (ETL)

Possible Data Destinations



Data Table

Extraction

Preparation

Predictive Model

Load Output

Data Warehouse

Credit One LLC

06

ADDITIONAL INSIGHTS

Other Ideas Based on Review of Data

Additional Insights of Data

- Extreme Outliers in Billing Statement Amount – ($334K), ($170K), and ($151K) balances (credits) in April and June 2005 – we will determine if these should be eliminated during model training and evaluation
- Ensure that binary indicator for Default Payment classifier is a String / Factor for classification model training in Python
- Consolidate additional indicators for Education – “0,4,5,6” should be consolidated into one indicator for simplification of model
- Determine viability of creating new attributes that can summarize multiple fields, such as a binary string indicator for any late payments during the six months – “0” any late payments or “1” no late payments