|  |
| --- |
| Gallogly College of Engineering, University of Oklahoma |
| Petabyte Pirates Project Report |
| Link to Checkpoint Review Presentation: <> |

|  |
| --- |
| ACS 5513 – Machine Learning Practice  Summer 2025 |

Contents

[Project Charter 1](#_Toc35593213)

[Business background 1](#_Toc35593214)

[Scope 1](#_Toc35593215)

[Personnel 1](#_Toc35593216)

[Metrics 1](#_Toc35593217)

[Plan 1](#_Toc35593218)

[Architecture 2](#_Toc35593219)

[Communication 2](#_Toc35593220)

[Data Report 3](#_Toc35593221)

[Data sources 3](#_Toc35593222)

[Raw Data Sources 3](#_Toc35593223)

[Processed Data 3](#_Toc35593224)

[Feature Sets 4](#_Toc35593225)

[Data dictionary 4](#_Toc35593226)

[Data report 5](#_Toc35593227)

[Software Architecture Document 6](#_Toc35593228)

[Model Report 7](#_Toc35593229)

[Analytic Approach 7](#_Toc35593230)

[Model Description 7](#_Toc35593231)

[Results (Model Performance) 7](#_Toc35593232)

[Model Understanding 7](#_Toc35593233)

[Conclusion and Discussions for Next Steps 7](#_Toc35593234)

[Exit Report 8](#_Toc35593235)

[Overview 8](#_Toc35593236)

[Business Domain 8](#_Toc35593237)

[Business Problem 8](#_Toc35593238)

[Data Processing 8](#_Toc35593239)

[Modeling, Validation 8](#_Toc35593240)

[Solution Architecture 8](#_Toc35593241)

[Benefits 8](#_Toc35593242)

[Company Benefit 8](#_Toc35593243)

[Customer Benefit 9](#_Toc35593244)

[Learnings 9](#_Toc35593245)

[Project Execution 9](#_Toc35593246)

[Data science / Engineering 9](#_Toc35593247)

[Domain 9](#_Toc35593248)

[Product 9](#_Toc35593249)

[What's unique about this project, specific challenges 9](#_Toc35593250)

[Links 9](#_Toc35593251)

[Next Steps 9](#_Toc35593252)

[Appendix 10](#_Toc35593253)

# [Project Charter](https://github.com/Azure/Azure-TDSP-ProjectTemplate/blob/master/Docs/Project/Charter.md)

## Business background

* Client: A Home Valuing Company Inc./University of Oklahoma
* Domain: Residential property valuation for appraisers, real-estate agents and home owners
* Business Problem:
  + Underwriters use rule-of-thumb comparable and simplistic regression tools that miss nonlinear interactions among houses and neighborhoods, age, and structural components.
  + This can create valuation errors that costs lenders and customers thousands of dollars in losses eventually leading to lost revenue.

## Scope

* The solution will be an end-to-end machine learning service that ingests raw sales and home-feature data using the Ames housing data, that trains a predictive model to show an on-demand price estimate.
* We will clean the data for over 3,000 records with 80+ features and train the models using logistic regression then tune them. We will also validate the model by holding out some data and cross-checking the model.
* How is it going to be consumed by the customer? Probably use a form to query data and then write back to response

## Personnel

* Who are on this project
  + - Project lead – Sean Miller
    - PM – Farhan Hassan
    - Data scientist(s) - Dewayne Hafenstein, Sean Miller
    - Account manager – Farhan Hassan
    - Data administrator – Sean
    - Business contact – Dewayne

## Metrics

* What are the qualitative objectives? (e.g. reduce user churn)
  + Compare what our app gives for suggested selling to what the property actually sold for +/- 5%
* What is a quantifiable metric (e.g. reduce the fraction of users with 4-week inactivity)?
* Quantify what improvement in the values of the metrics are useful for the customer scenario (e.g. reduce the fraction of users with 4-week inactivity by 20%)
* What is the baseline (current) value of the metric? (e.g. current fraction of users with 4-week inactivity = 60%)
* How will we measure the metric? (e.g. A/B test on a specified subset for a specified period; or comparison of performance after implementation to baseline) 50% baseline for the correlation

## Plan

* Phases (milestones), timeline, short description of what we'll do in each phase.

## Architecture

* Data: What data do we expect? Raw data in the customer data sources (e.g. on-prem files, SQL, on-prem Hadoop etc.)
* Data movement from on-prem to Azure using ADF or other data movement tools (Azcopy, EventHub etc.) to move either
  + all the data,
  + after some pre-aggregation on-prem,
  + Sampled data enough for modeling
* What tools and data storage/analytics resources will be used in the solution e.g.,
  + ASA for stream aggregation
  + HDI/Hive/R/Python for feature construction, aggregation and sampling
  + AzureML for modeling and web service operationalization
  + Host web portal on Heroku
  + Maybe Supabase?
  + Simple React + Typescript
* How will the score or operationalized web service(s) (RRS and/or BES) be consumed in the business workflow of the customer? If applicable, write down pseudo code for the APIs of the web service calls.
  + How will the customer use the model results to make decisions?
  + Data movement pipeline in production
  + Make a 1 slide diagram showing the end to end data flow and decision architecture
    - If there is a substantial change in the customer's business workflow, make a before/after diagram showing the data flow.

## Communication

* How will we keep in touch? Weekly meetings?
* Who are the contact persons on both sides?

# Data Report

## [Data sources](https://github.com/Azure/Azure-TDSP-ProjectTemplate/blob/master/Docs/Data_Report/Data%20Defintion.md#raw-data-sources)

### Raw Data Sources

| **Dataset Name** | **Original Location** | **Destination Location** | **Data Movement Tools / Scripts** | **Link to Report** |
| --- | --- | --- | --- | --- |
| Dataset 1 | Brief description of its original location | Brief description of its destination location | [script1.py](https://github.com/Azure/Azure-TDSP-ProjectTemplate/blob/master/Docs/Data_Report/link/to/python/script/file/in/Code) | [Dataset 1 Report](https://github.com/Azure/Azure-TDSP-ProjectTemplate/blob/master/Docs/Data_Report/link/to/report1) |
| Dataset 2 | Brief description of its original location | Brief description of its destination location | [script2.R](https://github.com/Azure/Azure-TDSP-ProjectTemplate/blob/master/Docs/Data_Report/link/to/R/script/file/in/Code) | [Dataset 2 Report](https://github.com/Azure/Azure-TDSP-ProjectTemplate/blob/master/Docs/Data_Report/link/to/report2) |

* Dataset1 summary. <Provide brief summary of the data, such as how to access the data. More detailed information should be in the Dataset1 Report.>
* Dataset2 summary. <Provide brief summary of the data, such as how to access the data. More detailed information should be in the Dataset2 Report.>

### Processed Data

| **Processed Dataset Name** | **Input Dataset(s)** | **Data Processing Tools/Scripts** | **Link to Report** |
| --- | --- | --- | --- |
| Processed Dataset 1 | [Dataset1](https://github.com/Azure/Azure-TDSP-ProjectTemplate/blob/master/Docs/Data_Report/link/to/dataset1/report), [Dataset2](https://github.com/Azure/Azure-TDSP-ProjectTemplate/blob/master/Docs/Data_Report/link/to/dataset2/report) | [Python\_Script1.py](https://github.com/Azure/Azure-TDSP-ProjectTemplate/blob/master/Docs/Data_Report/link/to/python/script/file/in/Code) | [Processed Dataset 1 Report](https://github.com/Azure/Azure-TDSP-ProjectTemplate/blob/master/Docs/Data_Report/link/to/report1) |
| Processed Dataset 2 | [Dataset2](https://github.com/Azure/Azure-TDSP-ProjectTemplate/blob/master/Docs/Data_Report/link/to/dataset2/report) | [script2.R](https://github.com/Azure/Azure-TDSP-ProjectTemplate/blob/master/Docs/Data_Report/link/to/R/script/file/in/Code) | [Processed Dataset 2 Report](https://github.com/Azure/Azure-TDSP-ProjectTemplate/blob/master/Docs/Data_Report/link/to/report2) |

* Processed Data1 summary. <Provide brief summary of the processed data, such as why you want to process data in this way. More detailed information about the processed data should be in the Processed Dataset1 Report.>
* Processed Data2 summary. <Provide brief summary of the processed data, such as why you want to process data in this way. More detailed information about the processed data should be in the Processed Dataset2 Report.>

### Feature Sets

| **Feature Set Name** | **Input Dataset(s)** | **Feature Engineering Tools/Scripts** | **Link to Report** |
| --- | --- | --- | --- |
| Feature Set 1 | [Dataset1](https://github.com/Azure/Azure-TDSP-ProjectTemplate/blob/master/Docs/Data_Report/link/to/dataset1/report), [Processed Dataset2](https://github.com/Azure/Azure-TDSP-ProjectTemplate/blob/master/Docs/Data_Report/link/to/dataset2/report) | [R\_Script2.R](https://github.com/Azure/Azure-TDSP-ProjectTemplate/blob/master/Docs/Data_Report/link/to/R/script/file/in/Code) | [Feature Set1 Report](https://github.com/Azure/Azure-TDSP-ProjectTemplate/blob/master/Docs/Data_Report/link/to/report1) |
| Feature Set 2 | [Processed Dataset2](https://github.com/Azure/Azure-TDSP-ProjectTemplate/blob/master/Docs/Data_Report/link/to/dataset2/report) | [SQL\_Script2.sql](https://github.com/Azure/Azure-TDSP-ProjectTemplate/blob/master/Docs/Data_Report/link/to/sql/script/file/in/Code) | [Feature Set2 Report](https://github.com/Azure/Azure-TDSP-ProjectTemplate/blob/master/Docs/Data_Report/link/to/report2) |

* Feature Set1 summary. <Provide detailed description of the feature set, such as the meaning of each feature. More detailed information about the feature set should be in the Feature Set1 Report.>
* Feature Set2 summary. <Provide detailed description of the feature set, such as the meaning of each feature. More detailed information about the feature set should be in the Feature Set2 Report.>

## [Data dictionary](https://github.com/Azure/Azure-TDSP-ProjectTemplate/tree/master/Docs/Data_Dictionaries)

Place to put data description documents, typically received from a client This is typically a field-level description of data files received.

This document provides the descriptions of the data that is provided by the client. If the client is providing data dictionaries in text (in emails or text files), directly copy them here, or have a snapshot of the text, and add it here as an image. If the client is providing data dictionaries in Excel worksheets, directly put the Excel files in this directory, and add a link to this Excel file.

If the client is providing you the data from a database-like data management system, you can also copy and paste the data schema (snapshot) here. If necessary, please also provide brief description of each column after the snapshot image, if such image does not have such information.

## [Data report](https://github.com/Azure/Azure-TDSP-ProjectTemplate/blob/master/Docs/Data_Report/DataSummaryReport.md#target-variable)

* General summary of the data
* Data quality summary
* Target variable
* Individual variables
* Variable ranking
* Relationship between explanatory variables and target variable

# Software Architecture Document

The solution architecture can be a diagram or description of your data pipeline that you use to run scoring or predictions on new data after you have built a model. It also contains the pipeline to retrain your model based on new data.

# Model Report

A report to provide details on a specific experiment (model) - possibly one of many

## Analytic Approach

* What is target definition
* What are inputs (description)
* What kind of model was built?

## Model Description

* Models and Parameters
  + Description or images of data flow graph
    - if AzureML, link to:
      * Training experiment
      * Scoring workflow
  + What learner(s) were used?
  + Learner hyper-parameters

## Results (Model Performance)

* ROC/Lift charts, AUC, R^2, MAPE as appropriate
* Performance graphs for parameters sweeps if applicable

## Model Understanding

* Variable Importance (significance)
* Insight Derived from the Model

## Conclusion and Discussions for Next Steps

* Conclusion
* Discussion on overfitting (if applicable)
* What other Features Can Be Generated from the Current Data
* What other Relevant Data Sources Are Available to Help the Modeling

# Exit Report

This is concise document that includes an overview of the entire project, including details of each stage and learning. If a section isn't applicable (e.g. project didn't include a ML model), simply mark that section as "Not applicable". Suggested length between 5-20 pages. Code should mostly be within code repository (not in this document).

Customer: <Enter Customer Name>

Team Members: <Enter team member' names. Please also enter relevant parties names, such as team lead, Account team, Business stakeholders, etc.>

## Overview

<Executive summary of entire solution, brief non-technical overview>

## Business Domain

<Industry, business domain of customer>

## Business Problem

<Business problem and exact use case(s), why it matters>

## Data Processing

<Schema of original datasets, how data was processed, final input data schema for model>

## Modeling, Validation

<Modeling techniques used, validation results, details of how validation conducted>

## Solution Architecture

<Architecture of the solution, describe clearly whether this was actually implemented or a proposed architecture. Include diagram and relevant details for reproducing similar architecture. Include details of why this architecture was chosen versus other architectures that were considered, if relevant>

## Benefits

### Company Benefit

(internal only. Double check if you want to share this with your customer)

<What did our company gain from this engagement? ROI, revenue, etc>

### Customer Benefit

What is the benefit (ROI, savings, productivity gains etc) for the customer? If just POC, what is estimated ROI? If exact metrics are not available, why does it have impact for the customer?>

## Learnings

### Project Execution

<Learnings around the customer engagement process>

### Data science / Engineering

<Learnings related to data science/engineering, tips/tricks, etc>

### Domain

<Learnings around the business domain, >

### Product

<Learnings around the products and services utilized in the solution >

### What's unique about this project, specific challenges

<Specific issues or setup, unique things, specific challenges that had to be addressed during the engagement and how that was accomplished>

### Links

<Links to published case studies, etc.; Link to git repository where all code sits>

### Next Steps

<Next steps. These should include milestones for follow-ups and who 'owns' this action. E.g. Post- Proof of Concept check-in on status on 12/1/2016 by X, monthly check-in meeting by Y, etc.>

# Appendix

<Other material that seems relevant – try to keep non-appendix to <20 pages but more details can be included in appendix if needed>