Group

MidTerm Project

Light House Labs Jan 2021

Table of Contents

[Executive Summary 0](#_Toc63407000)

[Objectives & Success Criteria 0](#_Toc63407001)

[Assumptions, Constraints/Dependencies, Requirements 0](#_Toc63407002)

[High-Level Risks 0](#_Toc63407003)

[Communication 1](#_Toc63407004)

[Daily Communication 1](#_Toc63407005)

[Project Meetings 1](#_Toc63407006)

[Daily Meeting 1](#_Toc63407007)

[Milestone Meetings 1](#_Toc63407008)

[Change Management 1](#_Toc63407009)

[Deliverables 1](#_Toc63407010)

[Quality 1](#_Toc63407011)

[Summary Milestones 1](#_Toc63407012)

[Schedule 2](#_Toc63407013)

[Stakeholders 2](#_Toc63407014)

[Project Team 2](#_Toc63407015)

# Executive Summary

This is the project management document for the Light House Labs Midterm Project for Jan 2021, Group. This document serves as a guide for how the project will be executed.

# Background

The goal is to predict arrival delays of commercial flights. Often, there isn't much airlines can do to avoid the delays, therefore, they play an important role in both profits and loss of the airlines. It is critical for airlines to estimate flight delays as accurate as possible because the results can be applied to both, improvements in customer satisfaction and income of airline agencies.

1. **Predicting flight delays:** In this project, you will practice the following topics
   * Supervised Learning
   * Feature Engineering
   * Sampling
   * JDBC connection to a database

### **Files**

* **exploratory\_analysis.ipynb**: this file contains 10 questions we need to answer during the data exploration phase. They will help us to understand the data and become familiar with different variables.
* **modeling.ipynb**: this file contains instructions for modeling part of the project. We recommend to split modeling tasks into more notebooks.
* **data\_description.md**: when you need to look for any information regarding specific attributes in the data this is the file to look in.
* **sample\_submission.csv**: this file is the example of how the submission of the results should look like.

### **Data**

We will be working with data from air travel industry. We will have four separate tables:

1. **flights**: The departure and arrival information about flights in US in years 2018 and 2019.
2. **fuel\_comsumption**: The fuel consumption of different airlines from years 2015-2019 aggregated per month.
3. **passengers**: The passenger totals on different routes from years 2015-2019 aggregated per month.
4. **flights\_test**: The departure and arrival information about flights in US in January 2020. This table will be used for evaluation. For submission, we are required to predict delays on flights from first 7 days of 2020 (1st of January - 7th of January).

Diagram

Description automatically generated

*“Feature engineering is the process of transforming****raw data****into****features****that better represent****the underlying problem****to****the predictive models****, resulting in improved****model accuracy****on****unseen data****.”*

*— Dr. Jason Brownlee*

From <[*https://towardsdatascience.com/understanding-feature-engineering-part-1-continuous-numeric-data-da4e47099a7b*](https://towardsdatascience.com/understanding-feature-engineering-part-1-continuous-numeric-data-da4e47099a7b)>

# Objectives & Success Criteria

|  |  |
| --- | --- |
| Objectives | Success Criteria |
| Data processing | Clean dataframe with relevant features |
| Modelling | Many iterative approaches tried, with varying R2 reported |
| Results | Highest R2 relative to base model |
| Presentation | All team members feel represented and have a chance to contribute |
| Github Repository | All team members have access, content is clean and concise with comments/markdown giving context |
| Complete project | All team members feel represented and have a chance to contribute |

# Assumptions, Constraints/Dependencies, Requirements

|  |
| --- |
| Assumptions |
| Feature engineering will be 80% of project time |
| Mentor help available to get through stuck points |

|  |
| --- |
| Constraints/Dependencies |
| Team members availability for group work |

|  |
| --- |
| Requirements |
| Data |
| Computer hardware and software (Google Colab) |

# High-Level Risks

|  |  |  |
| --- | --- | --- |
| Risk | Level | Mitigation |
| Communication breakdown | High | Daily meetings  Project Plan  Change management section  Discord table  Slack group |
| Project documentation incomplete | High | Remain vigilant in good behaviour from the start  Spend time each day to work on keeping it clean/clear  Put deadline 1 week after Project Close to complete final |
| Project runs out of time | Medium | This project plan helps identify key work packages and timelines so that decisions to move forward when necessary can be made |

# Communication

## Daily Communication

* Discord chat table
* Slack group chat

## Project Meetings

### Daily Meeting

A daily meeting should occur to touch base with all team members and discuss:

1. Progress/Status
2. Risks/Challenges
3. Next work/Timeline updates

This will serve as a predictable opportunity to discuss the project at a high level and connect additional help when risks are identified.

**Proposed time: Daily 10:00 - 10:30am PST**

### Milestone Meetings

Ahead of major milestones (Presentation, Report, Final Code/Git upload) a meeting may be required to align the team on the deliverable.

**Proposed: Add additional time to daily meeting 10:30am – 12:00pm PST (if necessary)**

# Change Management

When the need to change an agreed upon component of the project arises, notification of the team with rationalization and written record of the change should occur. **Recommendation is to do this by email.**

# Deliverables

* Project Plan
* Timeline
* Work Packages/[WBS](#_Work_Breakdown_Structure)
* Presentation
* Report

## Work Breakdown Structure

|  |  |  |
| --- | --- | --- |
| Tasks | Est Completion Date | Notes, Dependencies |
| Data Exploration | Sun Feb 7 | What do we have? |
| Data Cleaning | Sun Feb 7 | Remove nulls, empty columns, encode |
| Baseline Model (MVP) | Mon Feb 8 | Obtain baseline accuracy/R2 |
| Iterate models | Tues Feb 9 – Wed Feb 10 | Obtain model with highest accuracy   * linear / logistic / multinomial logistic regression * Naive Bayes * Random Forest * ~~SVM~~ * XGBoost * The ensemble of your own choice   Classification versus regressor for each model (late/not – if yes, then range). Predict 7 days in advance. |
| Select best model, tune (4-5 hours) | Thur Feb 11 | Model with highest accuracy from iteration should be tuned to find accuracy > 95% |
| Run test data, evaluate outcome | Thur Feb 11 | Using 2020 data, run the model to make the predictions |
| Make business observations | Thur Feb 11 | How does this model help the business? |
| Create presentation | Fri Feb 12 | As per the guidelines |
| Submit file | Fri Feb 12 | As per the guidelines |

# Quality

How will we decide what good looks like?

* ~~Accuracy >95% (Anurag, 2021)~~
* R2 performance increases over iteration of models

How will we know when we are complete?

* ~~When we get a model that can predict with 95% accuracy, we will run our 2020 data~~
* When we get a model of accepted R2, we will run our 2020 data
* We will use the output of this to predict delays and make business decisions

# Summary Milestones

|  |  |  |
| --- | --- | --- |
| Date | Item | Notes |
| Fri Feb 5 | Project Kick Off Meeting |  |
| Sun Feb 7 | Data Cleaning/Exploration completed |  |
| Mon Feb 8 | MVP created | Initial accuracy/R2 recorded |
| Thur Feb 11 | Finalize model  Run 2020 test data  Finish Presentation |  |
| Fri Feb 12 | Project Presentation  Submit final file  Project Close |  |

# Schedule

|  |  |
| --- | --- |
| Date | Item |
| Fri Feb 5 | Project Kick Off Meeting  Identify problem statement  Start to identify Project work packages |
| Sat Feb 6 | Project Meeting |
| Sun Feb 7 | Project Meeting |
| Mon Feb 8 | Project Meeting |
| Tues Feb 9 | Project Meeting |
| Wed Feb 10 | Project Meeting  Choose model to tune |
| Thur Feb 11 | Project Meeting  Complete model  Run test data  Finish Project Presentation |
| Fri Feb 12 | Project Meeting  Project Presentation  Project Close |

# Stakeholders

|  |  |  |
| --- | --- | --- |
| Name | Position | How are they utilized |
| Mentors | Mentor | Advice on project execution, model evaluation, output/presentation |
| Light House Labs | School | Provide assessment of project |
| LHL Cohort | Classmates | Advice and comparison of approaches, shared learnings |

# Project Team

|  |  |  |
| --- | --- | --- |
| Name | Phone Number | Email |
| Stacey McLennan-Waldal |  |  |
| Devika Sunil |  |  |