NHL Win Prediction

Team No.	5509-9
Team Name	NHL Pros
GitHub Repository	NHL-Game

Product Manager	Anqi Chen	angelach99
Business Analyst	Matthew Buttler Ives	matt-buttlerives
Data Analyst	Mesaye Bahiru	mesaye3
	Kexin Chen	KEX9027
	Rebecca Mukena Yumba	beccarem
Data Scientist	Betty Au	bettyau
	Claire Cai	clairecai97

Agenda

1. Business Case Overview
2. Data Acquisition
3. Data Exploration
4. Data Pre-Processing
5. Modelling
6. Causal Analysis
7. Result Analysis

Business Case Overview

1. **Many Different Types of Betting for the NHL -** Moneyline, Puckline, Over/Under

2. **Live Betting -** Can we develop a model to make accurate predictions on which team will win the game based on the first period of play?

3. **Proof of Value -** Random Guessing vs. Point Spread

4. **Notoriously Difficult -** Best models' performance ranges from 50-70% accuracy. Excellent model now is 62% (Forecheck).

Data Acquisition & Transformation

I

Merge Data Files,
Select Relevant Columns to Add

(e.g., teams away or at home,
time on ice)

II

Final Data Frame

- **Games:** date/time, venue, stats
- **Teams:** away or at home, stats
- **Target:** win or lose

III

	game_id	team_id	HoA	won	settled_in	head_coach
0	2016020045	4	away	False	REG	Dave Hakstol
1	2016020045	16	home	True	REG	Joel Quenneville
2	2017020812	24	away	True	OT	Randy Carlyle
3	2017020812	7	home	False	OT	Phil Housley

Data Exploration

Dataset Size	7.2 MB
Dataset Shape	52,610 Rows & 25 Columns
Variable Type	2 IDs + 17 Numeric + 6 Categorical
Missing Values	Percentage of Face Off Wins: 22,148

Steps to Take:

Data Distribution

- Statistics Summary (mean, standard deviation, median, minimum, maximum, 25-percentile, 75- percentile)
- Value Count
- Histogram
- Box Plot
- Scatter Plot

Correlation

- Correlation Matrix
- Heatmap

IDs

gameID, teamID

Categorical Variables

hoa, won, settledIn, headCoach, startRinkSide, goalieReplacement

Numeric Variables

pim, powerPlayOpportunities, powerPlayGoals, faceOffWinPercentage, shots, goals, takeaways, hits, blocked, shotsGiveaways, missedShots, penalties, wonFaceOffs, timeOnIce, evenTimeOnIce, shortHandedTimeOnIce, powerPlayTimeOnIce

Data Preprocessing

Missing Values

Numeric Variables

faceOffWinPercentage, timeOnIce, evenTimeOnIce, shortHandedTimeOnIce, powerPlayTimeOnIce, pim, powerPlayOpportunities, powerPlayGoals, teamID

Imputation via Median Approach

Categorical Variables

startRinkSide, goalieReplacement

One-Hot Encoding

Multicollinearity

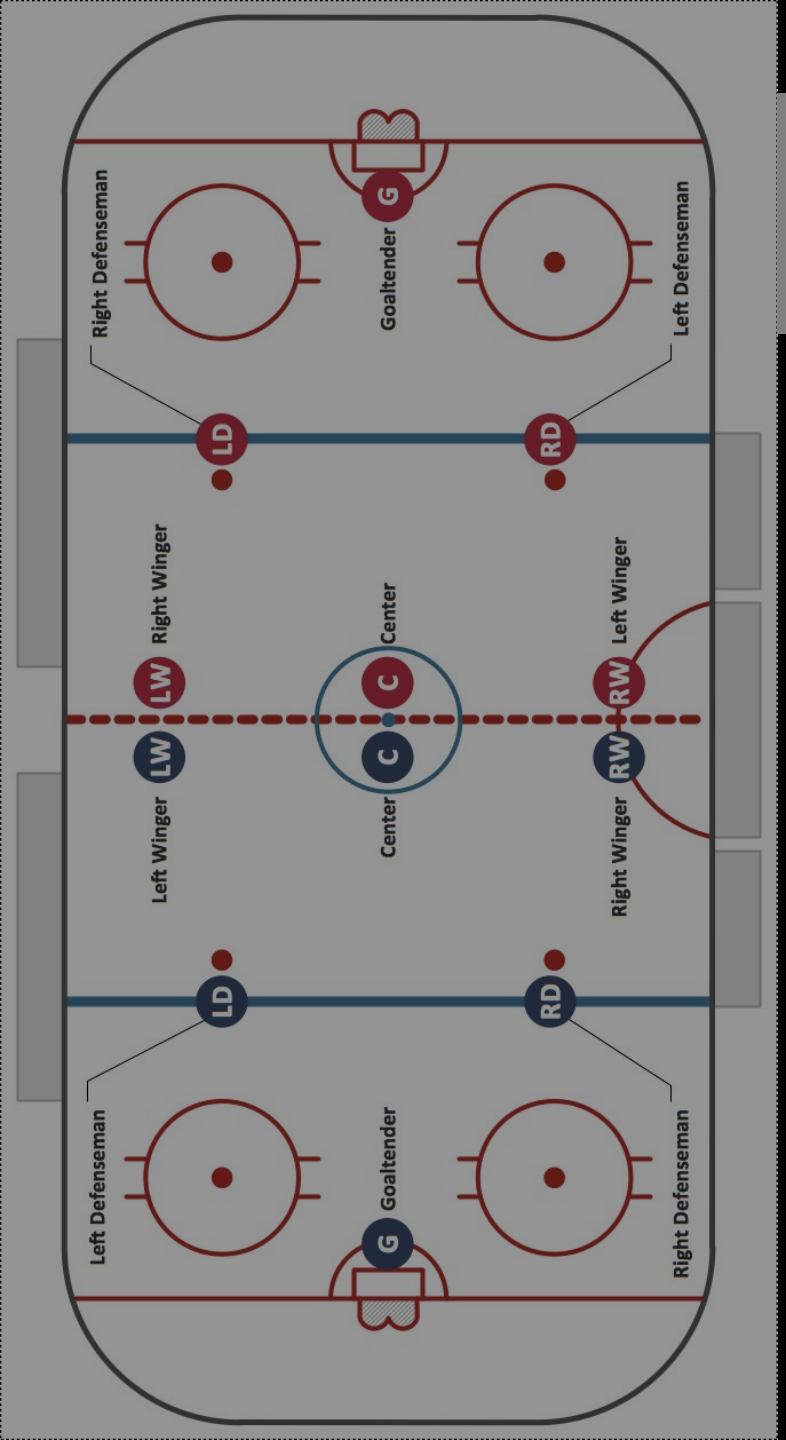
powerPlayTimeOnIce	*powerPlayOpportunities*
gameID	*wonFaceOff*
shots	*wonFaceOff*

Feature Selection Process

settledIn_tbc: the least important predictor

Outliers' Detection

most outlier variables located in:
timeOnIce,
evenTimeOnIce



Modelling

Business Objectives

Use the first period game data to predict the winning team

Business Goals

To provide additional insights and achieve an accuracy at or above 53%

Models

- Logistic Regression
- ANN
- Random Forest
- Gradient Boosting

Causal Inference Analysis

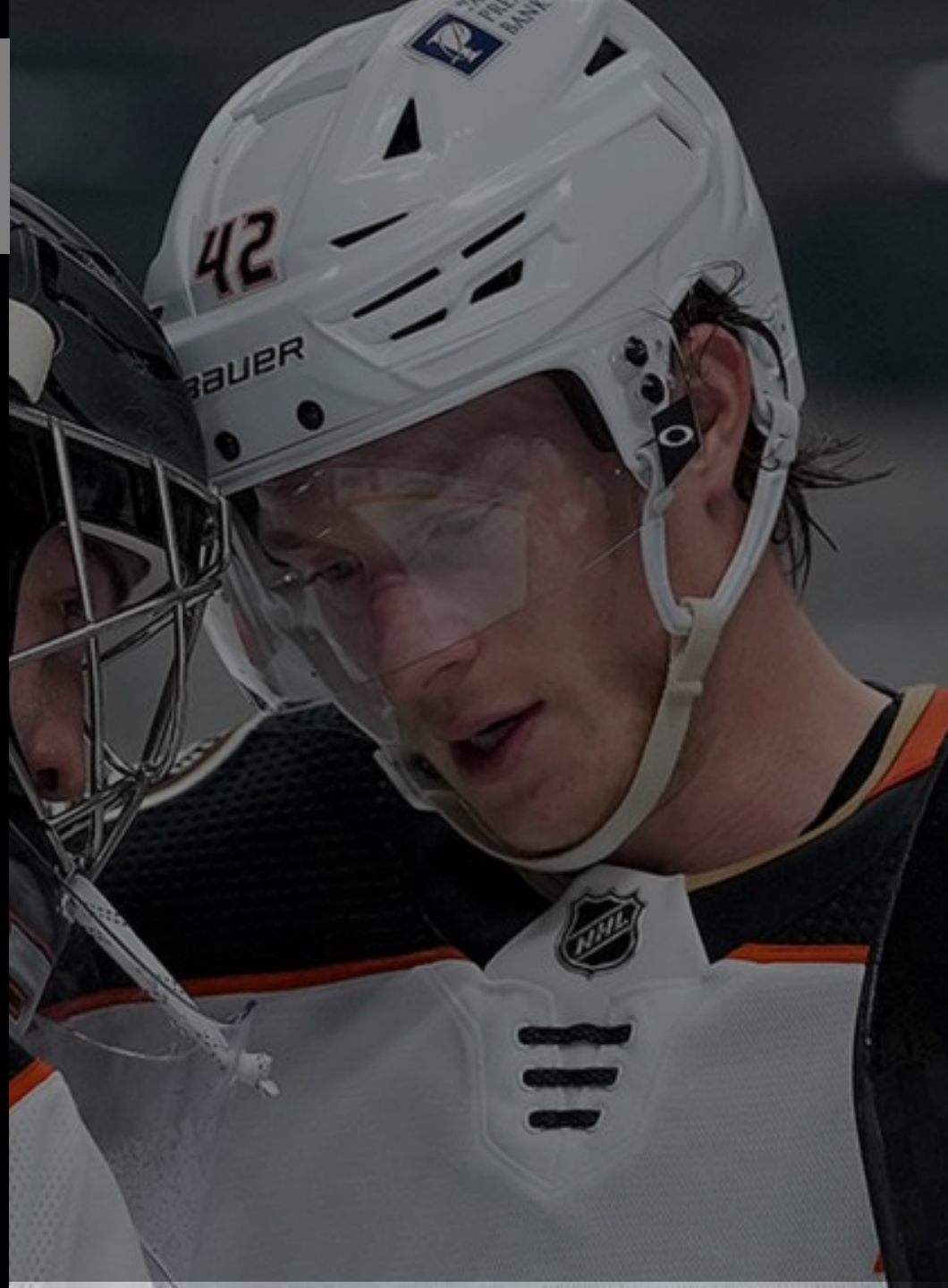
Objective - to examine the difference in win by considering the treatment of whether the game happened at home or away, and to study the feature importance of explanatory variables

Outcome Variable - whether the team win the game or not

Treatment Variable - whether the team is at home or away

Explanatory Variables - 19 remaining variables

Modelling - use *CausalML* package to run various causal analysis classification models with different learners



Results Analysis

ML Models

- Help to make accurate predictions on sports betting over randomly guessing the outcome
- Investigate the important factors that contribute to the outcome of a hockey game

Causal Inference

- Decision-makers can quantify the influence of home games and act accordingly on sports betting

Limitation

- Models with accuracy levels ranging from 50% to 70%
- Best models typically only perform well for one season

Thank You!