**Cox-Lauf – MSDS696 Practicum II Project Proposal**

**1. Name, Contact info (e.g. email/phone):**

Elan (“Lonny”) Cox-Lauf, [ecoxlauf@regis.edu](mailto:ecoxlauf@regis.edu), [lonnyjcox@gmail.com](mailto:lonnyjcox@gmail.com), (303) 522-0748

**2. Title of the project:**

NHL Expected Goal Predictions using Shot Statistics and Machine Learning

**3. High level description of the project: what question or problem are you addressing?**

In this project I will be addressing the feasibility of using machine learning models to predict whether a shot in the NHL will result in a goal. To do this, a statistical measure/probability called “expected goal” or “xG” will be utilized to make these predictions. One question that I would aim to answer in this project is: which shot-related statistics (i.e., shot speed in MPH, location on the ice (x/y coords), the circumstances of the play (off rebound or pass, unassisted basic shot, etc.), time elapsed since the last shot, shot type (wrist shot, slap shot, …), and several others) contribute the most heavily toward the expected goal metric. Having answers to questions like this one will help coaches and franchises determine what to focus on in games to increase the likelihood of winning.

**4. What type of data science task is it?**

This is a prediction/feature selection problem using supervised learning.

**5. Data: Brief description of data. How big do you expect the data will be? Is amount of**

**your data too big or too small? If you're web-scraping or collecting data, how long do you**

**expect to collect the data?**

I have found several potential web sources, such as <https://hockey-scraper.readthedocs.io/en/latest/>, <https://records.nhl.com/records/playoff-team-records/> and <https://www.hockey-reference.com/>. These sources have different historical statistics for all teams, and some of the statistics go back until the first Stanley Cup Finals in 1917-1918. The first link is a web-scraper that extracts play-by-play data from 2007 and on to the present season. There will be sufficient data since there are about 1300 games per season total, and about 60 shots taken per game.

**6. How will you analyze the data? What machine learning methods do you plan to use,**

**and/or what business intelligence aspect do you plan on incorporating?**

After EDA and feature selection techniques to remove features that are not relevant, I plan to use several types of supervised machine learning models to predict expected goals. Some of these models I intent to try include: decision tree, binary logistic regression (binary since each team will either 0: no goal, or 1: goal), and neural network. Others may be attempted to as I learn more and progress through the project.

**7. Describe any anticipated difficulties and problems. Discuss how you may overcome the**

**problems.**

I anticipate that there will be difficulty with the feature selection since there are so many features that are recorded in NHL statistics for all the shots taken. It will be a challenge to identify which features should remain and which should be thrown out prior to feeding them to the models I will train. It will also be challenging to fill in any missing data/stats for certain shots taken in the NHL, as well as making sure the data is normalized and/or converted to useable format. Also, some of the features will be categorical (i.e., shot type) while other may be quantified (i.e., shot speed and angle, etc.).

**8. Suggest a timeline for the project. This should be a weekly breakdown of what you plan**

**on doing each week.**

**Week 2:** Data scraping, cleaning, and formatting into useable format (likely dataframe)

**Week 3:** Feature selection techniques to determine key team statistics as well as nonessential statistics for the models. Perform training and test data split.

**Week 4:** Train and evaluate binary logistic regression model (or SVM)

**Week 5:** Train and evaluate decision tree model (or perhaps Random Forest)

**Week 6:** Train and evaluate neural network model

**Week 7:** Compare results, choose the best, and refine to improve the metrics on the test set. Begin creating visualizations for presentation, etc.

**Week 8:** Create, practice, and record presentation using PowerPoint

**9. Create GitHub repository for your Practicum project. Add this proposal, begin a ReadMe**

**document, and begin adding your data to your repository. Add a link to your GitHub**

**repository to this document**

**GitHub Link:** <https://github.com/LonnyCox55/MSDS696_NHLExpectedGoalPrediction>