





# Pittsburgh Penguins Capstone Final Presentation



Josh Dixon,  
Kayko Ohkawa,  
Spencer Segal



# Goal



Create a model that is able to estimate salaries for minor league hockey players

# Why is predicting salaries important?

---

- Long term financial planning
- Can identify players that are being overpaid vs market value
- Identify cheap talent

# Data Acquisition

---

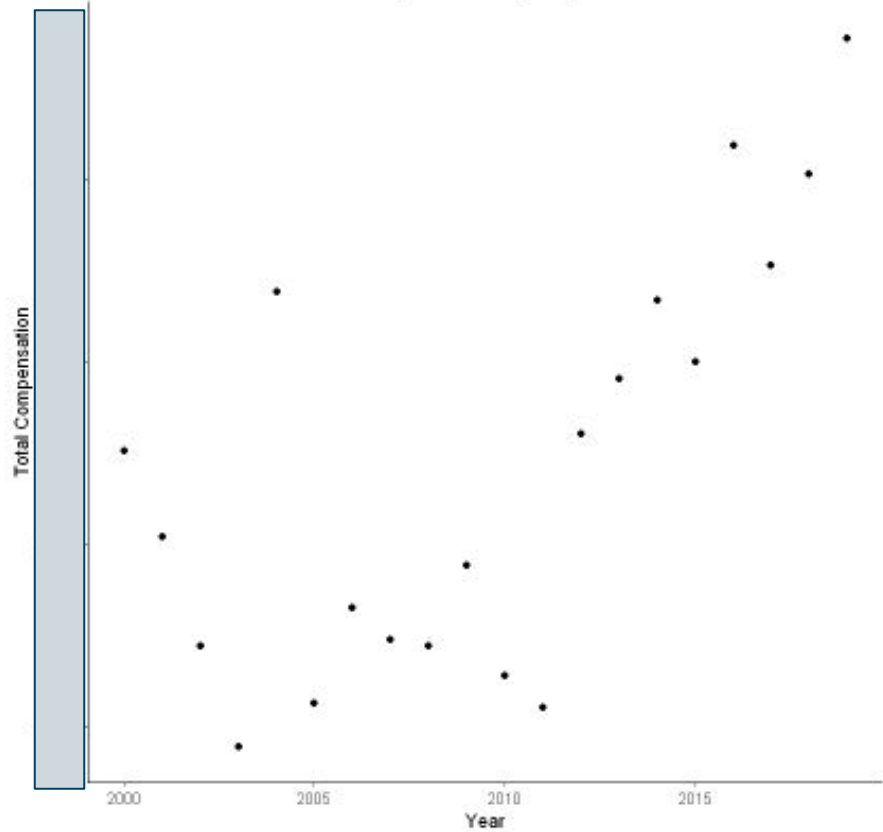
- Web Scraping
  - From CapFriendly and Elite Prospects
- Manual Data Entry
  - From Word documents
- Data Mining
  - From Excel documents

# Data Cleaning

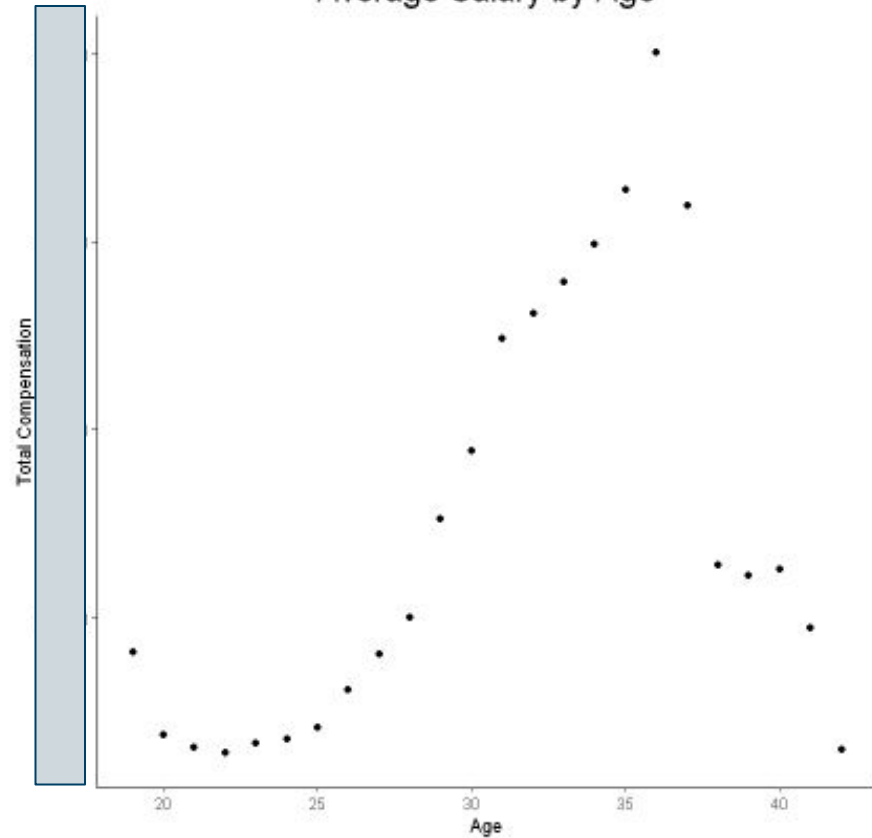
---

- Matching Player Names - Fuzzy Matching
  - And to a lesser extent, team names (could do manually)
- Inflation/Conversion Rates
- Signing Bonuses - Regular Expressions
- Total Compensation
- Per Game Stats
- Combining Data Sets

### Average Salary by Year

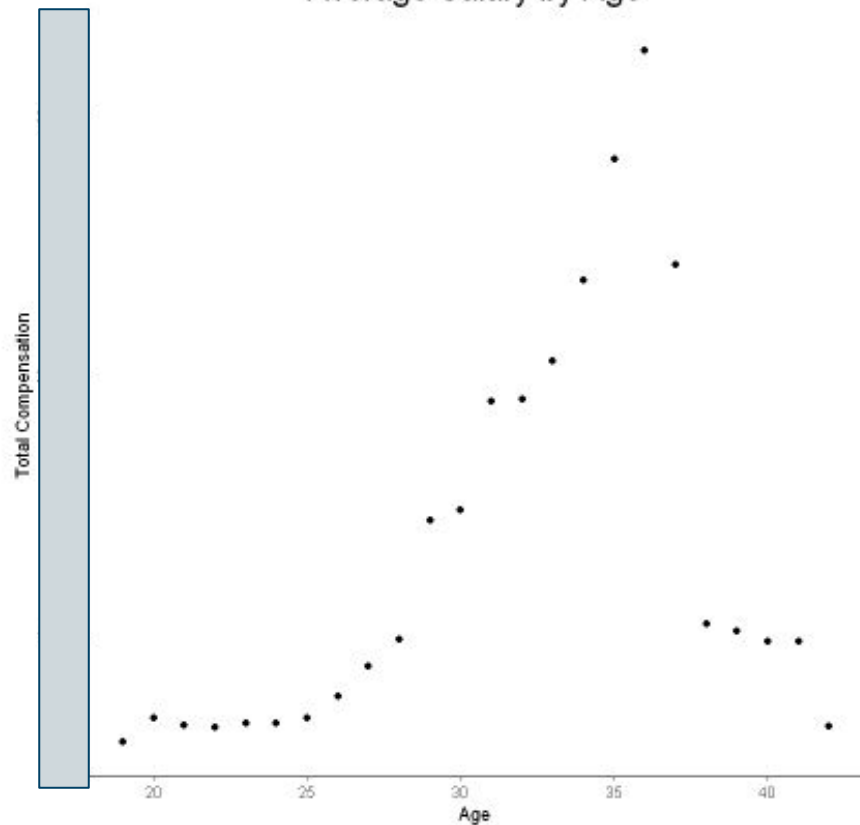


### Average Salary by Age



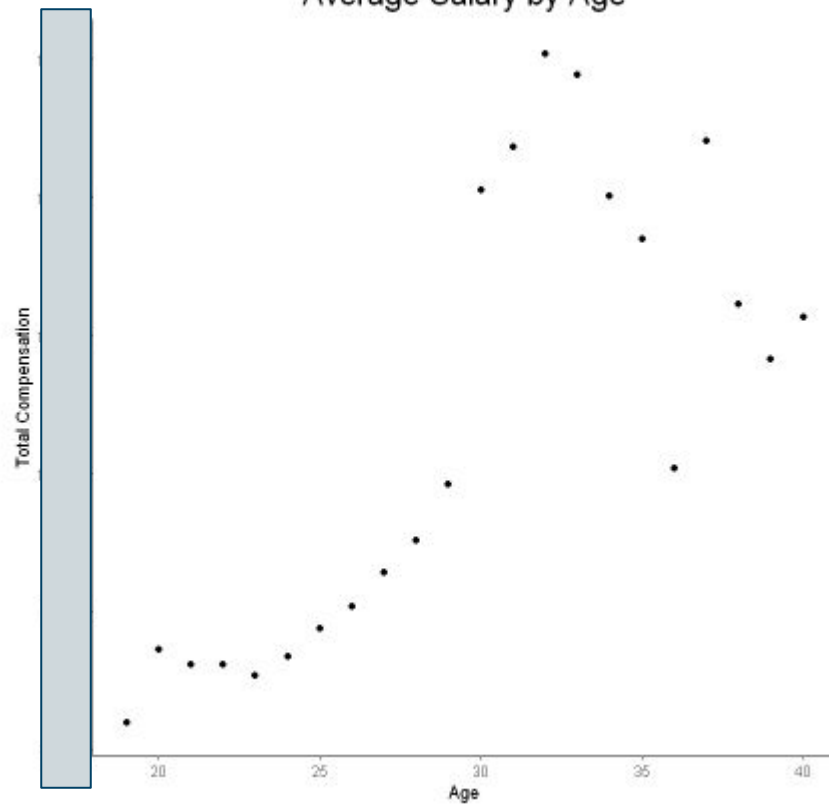
# Forwards

Average Salary by Age



# Defense

Average Salary by Age



# Linear Model Creation

---

## Specifications:

Dropped 2004 year stats - Outlier

Interacting Year with Player Statistics

Split the data into Forwards and Defense

Squared/cubed Age - accounts for non-linear relationship

Limit looking back to 1 year - 2, 3, 4 years back did not impact Goodness of fit



# Other Models: Forwards

---

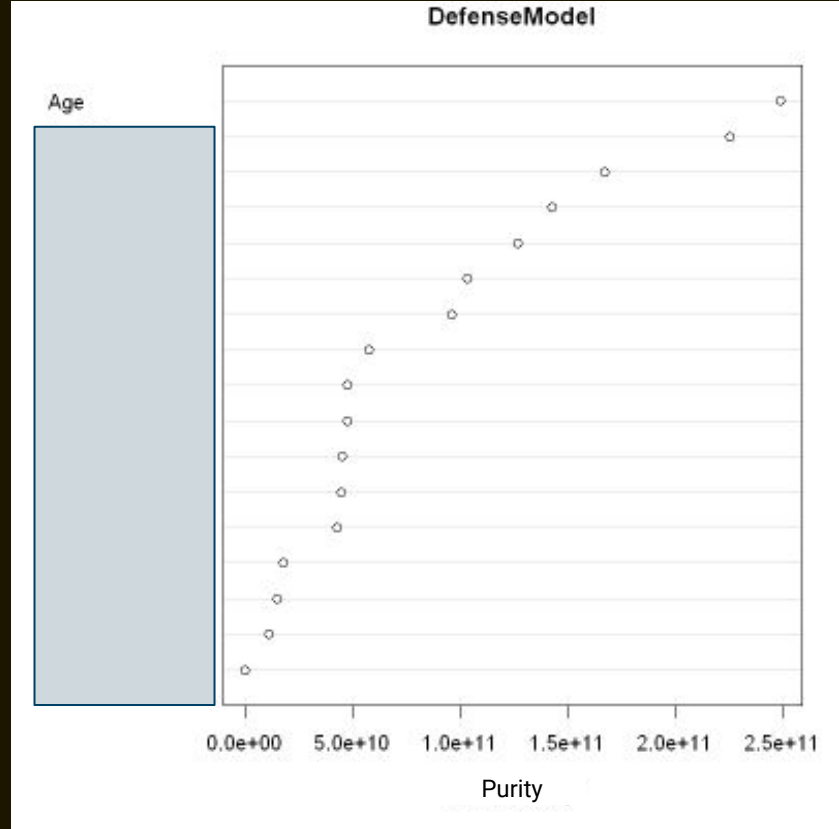
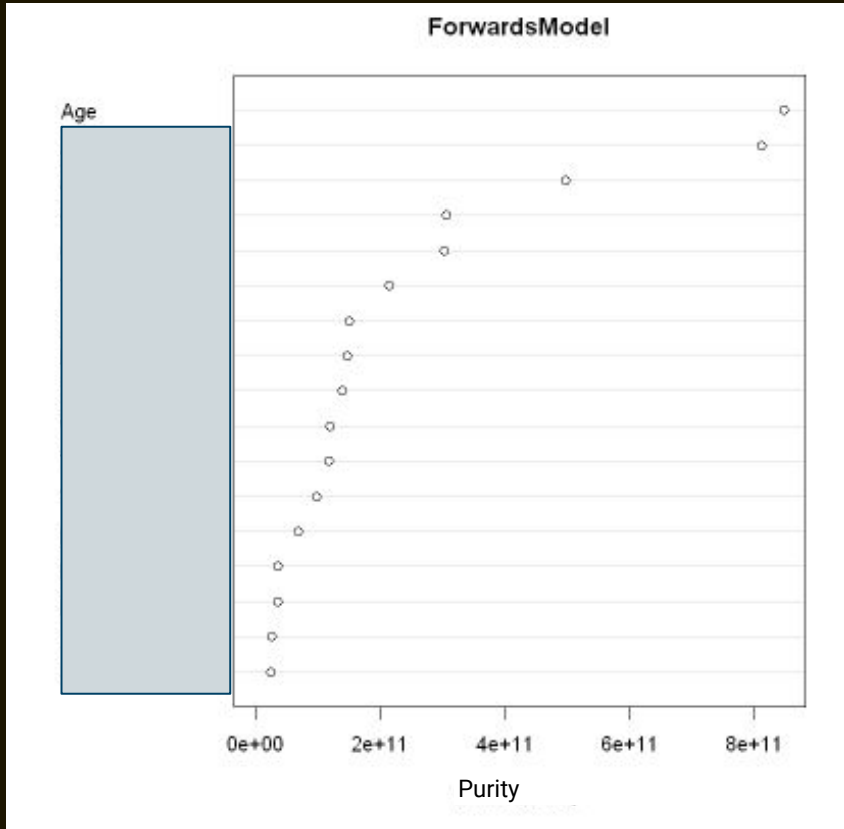
Statistics	<u>MSPE</u>	<u>SQRT MSPE</u>	<u>MAE</u>	<u>MAE Percent Error</u>	<u>MEAN</u>
<u>Linear</u>	1,745,199,375.57	\$41,774.05	\$24,700.09	0.275	\$88,442.78
<u>Random Forest</u>	1,612,272,582.077	\$40,150.69	\$22,120.83	0.229	\$88,424.36
<u>XGBoost</u>	2,255,512,758.851	\$47,488.20	\$24,553.608	0.251	\$87,408.78

# Other Models: Defense

---

Statistics	<u>MSPE</u>	<u>SQRT MSPE</u>	<u>MAE</u>	<u>MAE Percent Error</u>	<u>MEAN</u>
<u>Linear</u>	1,449,450,725.65	\$38,067.29	\$22,869.35	0.263	\$83,352.71
<u>Random Forest</u>	1,279,101,535.799	\$35,762.38	\$20,228.29	0.227	\$83,365.71
<u>XGBoost</u>	1,703,521,123.516	\$41,268.06	\$22,554.31	0.248	\$82,524.52

# Best Model: Random Forest (Variable Importance)



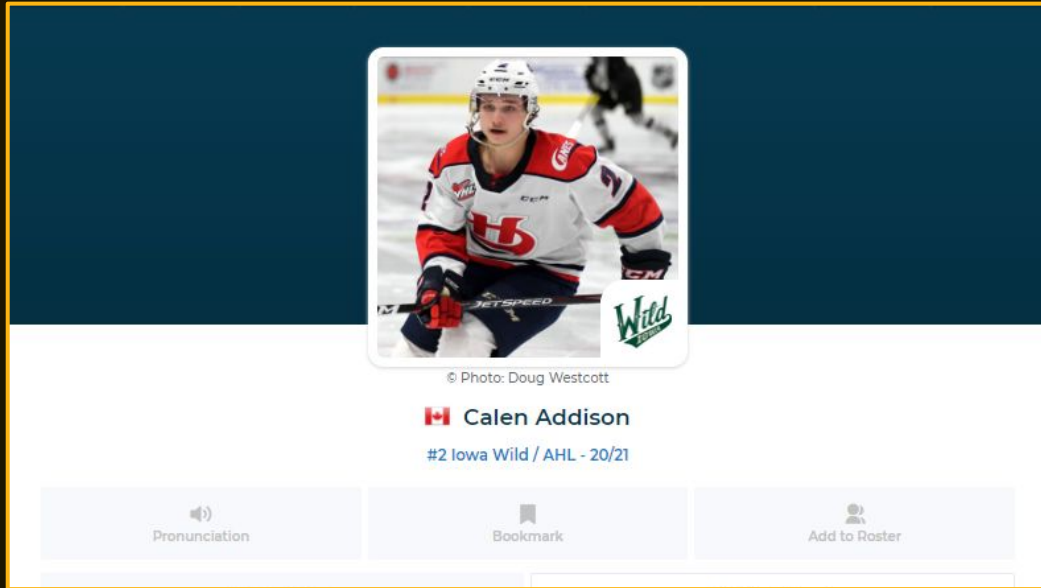
# Random Forest Comparison to NHL Models

---

Statistics	Our Random Forest Forwards Model	Our Random Forest Defense Model	Matt Cane's Updated NHL Model
Average Mean Absolute Error	\$22,121	\$20,228	\$429,000
Average Mean % Error	22.9%	22.7%	28.5%

# Interactive Interface

Demo: Provide an EP profile link >> Get the recommended AHL contract salary



A screenshot of an athlete profile page for Calen Addison. The profile features a photo of the player in a white and red hockey jersey with the number 2, playing for the Iowa Wild. Below the photo, the text reads "© Photo: Doug Westcott", "Canada Calen Addison", and "#2 Iowa Wild / AHL - 20/21". At the bottom of the profile, there are three interactive buttons: "Pronunciation", "Bookmark", and "Add to Roster".

>>

```
#####  
Name:      Calen Addison  
Estimate:  $76,466.7  
#####
```

# Model Limitations

---

If no previous-year AHL stats are available, predictions will be less precise

Only relevant for AHL-Only contract salary predictions

Still a fairly large MSE

Does not account for experience

Thank you!