Data Science Summer 2018 Internship Cases

GUSTAVO OLIVEIRA – SYRACUSE UNIVERSITY – ISCHOOL

APPLIED DATA SCIENCE – MASTERS PROGRAM

CONTACT: GOLIVEIR@SYR.EDU - 302-727-1599

Confidential

Industry

- One of the three largest container shipping companies in the world.
- LATAM regional office
- Large amount of non-integrated data
  - Large amount of exploratory analysis needed (not enough time to explore)
  - For all the projects, data needed to be cleaned and contextualized

Projects for the Internship

- 7 week internship projects:
  - Define a Tier Segmentation for clients.
    - Be able to price with more precision and speed
    - Have a hierarchical understanding and uniformity of pricing throughout the company
  - Develop Volume and Revenue Forecast Models for 1-2-3 weeks ahead
    - Today all reporting is explanatory and actions are reactive.
    - Idea is to be able to act in advance and maintain a high volume shipping rate.
- Guidelines of what was expected as results was provided. Tools and methods were suggested and developed by intern.
- Intern was also requested to provide recommendations to improve global data manipulation and integration throughout the office.

TIER SEGMENTATION

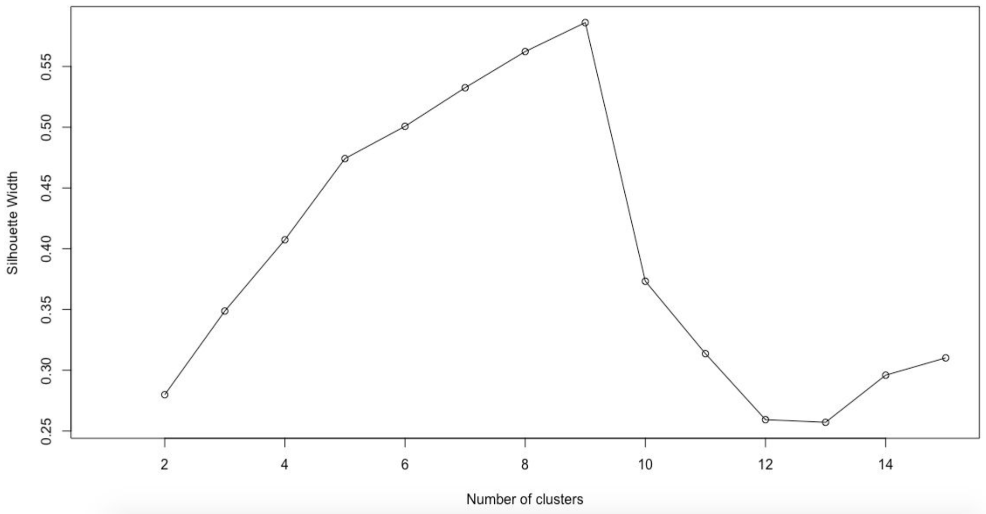
- Solution Proposed:
    - Use Cluster Analysis to define the segments and assist the Trade Directors to create the Tiers.
    - Proposition was very well accepted.
- Method used:
    - K-means cluster analysis with Gower's Distance
    - K-means – you pre define the number of clusters for the model by the K parameter.
        - In this case, we calculated from k=2 to k=15/20/40/100 and used the Silhouette Index to identify the optimal number of clusters.

TIER SEGMENTATION

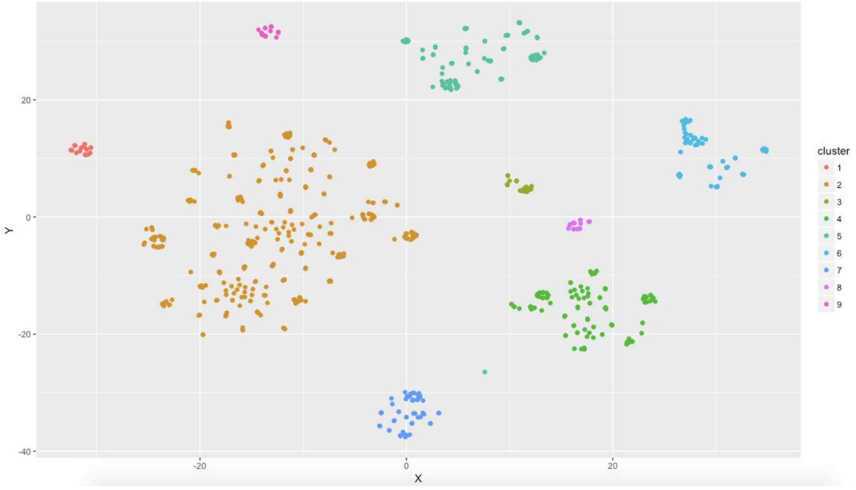
- Method used:
  - After significant amount of data analysis and specialist input we decided to use 3 variables for the cluster analysis. Two Categorical and one numerical variable, hence the use of Gowers Distance, which accounts for mixed type of variables.
  - With specialist input and many trials later it was decided to use weights on variables to give more strength to variables that experienced showed were more relevant to the business.
  - Once the clusters were created, specialists reviewed for validation.
  - The clusters were all used for Client Tier segmentation definition.

Some Results: Trade 1

Silhouette for best number of Clusters

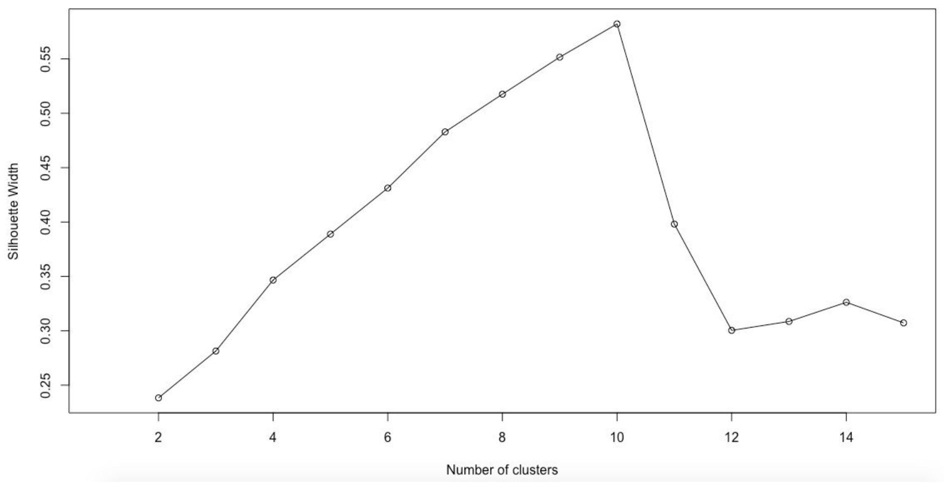


Clusters

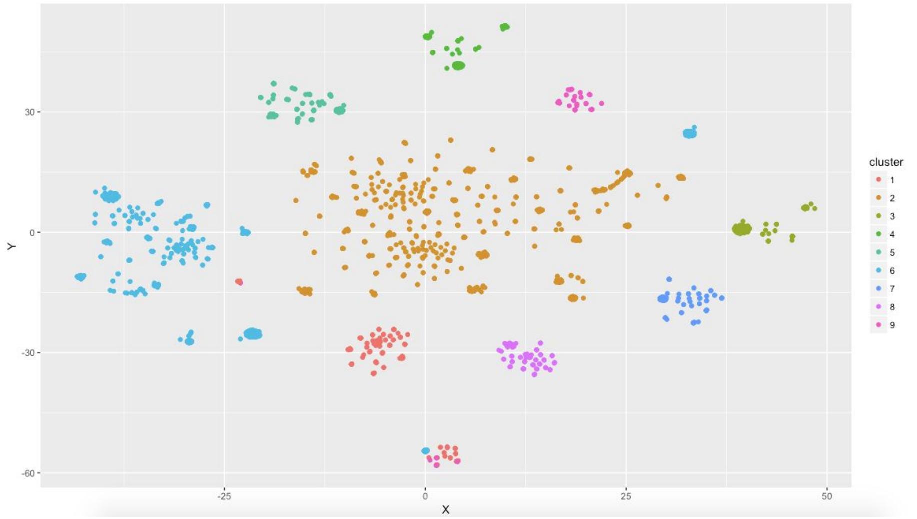


Some Results: Trade 2

Silhouette for best number of Clusters



Clusters



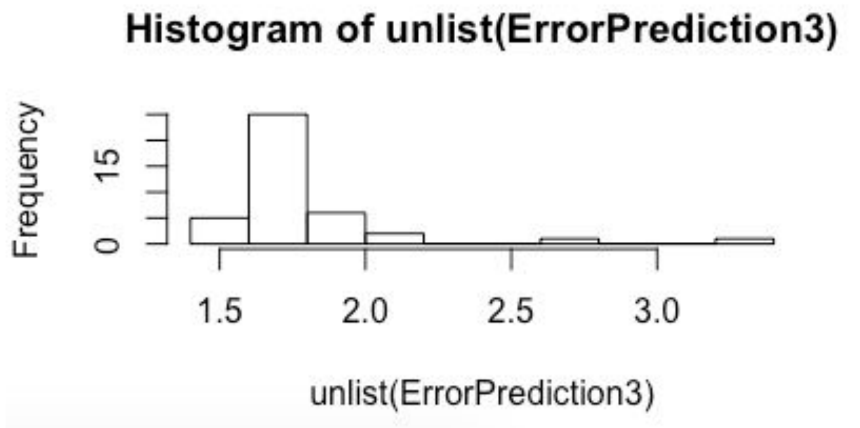
Confidential

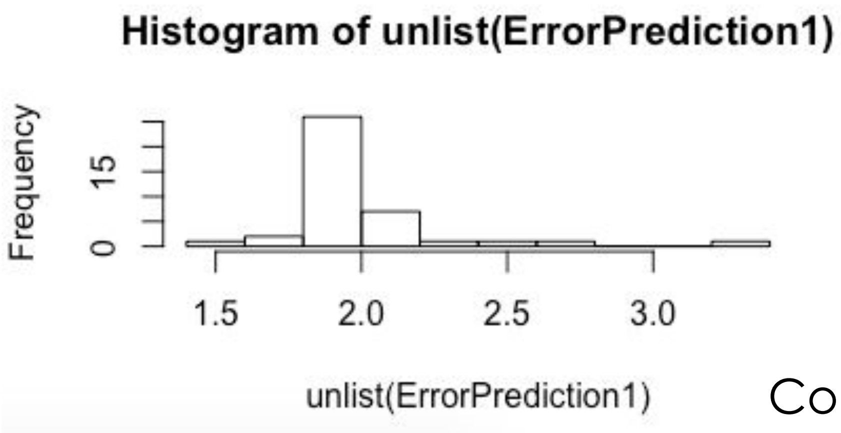
Volume and Revenue Forecast

- Data:
  - Forecast Volume and revenue per shipping voyage:
    - Revenue data not available in the breakdown needed immediately
    - Volume Data needed a lot of pre-processing to run the models.
    - Volume model can be extrapolated to forecast revenue.
- More than 1000 voyages, so efforts were concentrated on the most relevant for the Regional Office:
  - Modelling is an ongoing activity, so guidelines were defined for model definition so the work could be continued once the internship was over.

Volume and Revenue Forecast

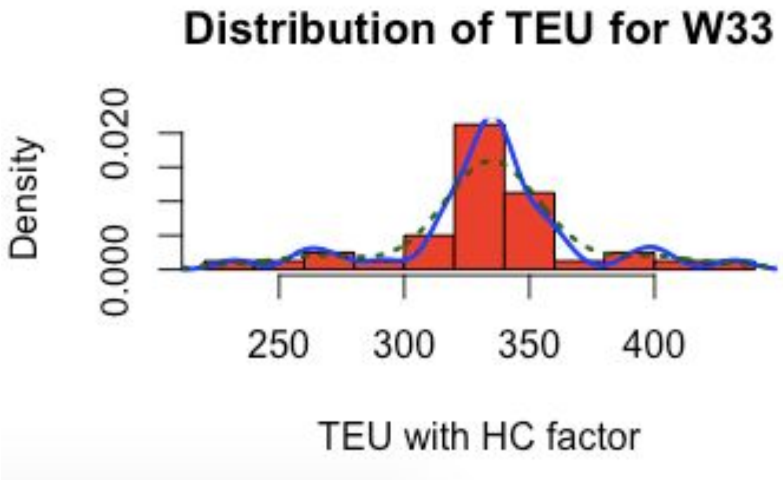
- Models Used:
  - After attempts with different models, RNN (Recurrent Neural Network), specifically Long Short-Term Model (LSTM) worked the best.
  - For cross validation and result confidence interval we used Monte Carlo Simulation on top of the neural net.
  - Error Analysis (% of error per simulation):



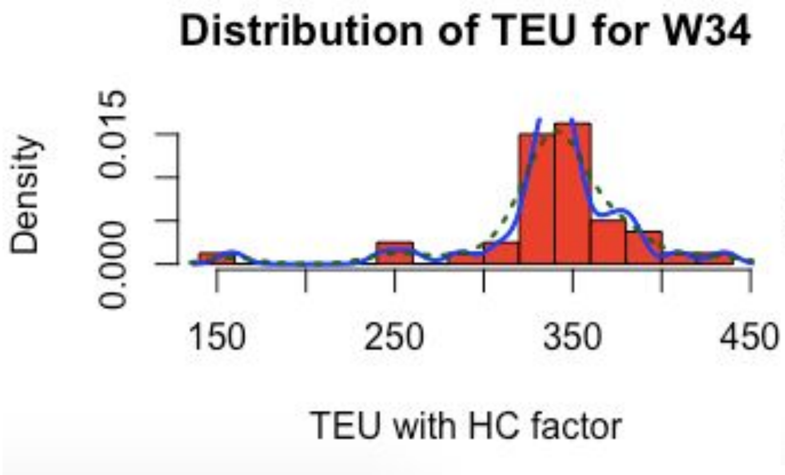


Volume Results

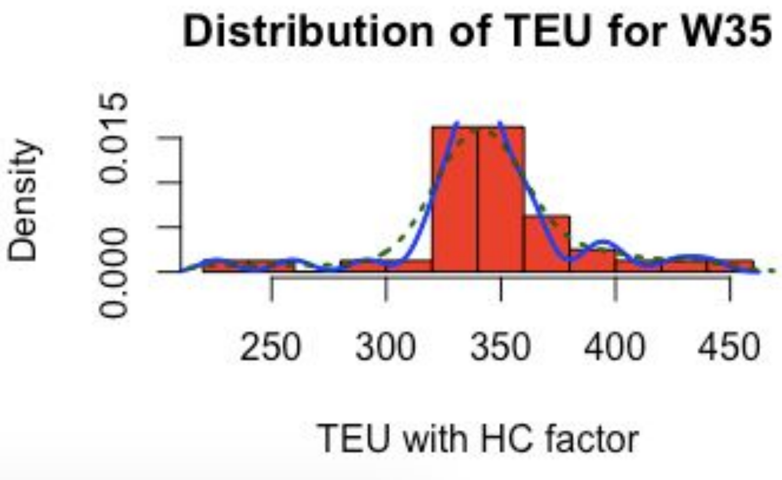
- **With a distribution/density graph you have a range and probability of forecast:**



Week 33



Week 34



Week 35

Summary

- Techniques used:
  - Cluster Analysis
    - K-means, Gower Distance
  - Recurring Neural Networks
    - Long Short-Term Model
  - Monte Carlo Simulation
  - Confidence interval
- Tools used
  - Excel VBA
  - RStudio

- Provided recommendations to improve overall data management, such as:
  - Pull data management from the IT department and create an independent Data Management and Analytics department (Responding directly to the SVP/CEO)
  - Guarantee data integrity across the company (all areas should be using the same correct data)
  - Integrate and develop automated tools for efficient data manipulation.

Intern Contact Info

Gustavo Oliveira – Syracuse University – iSchool

APPLIED DATA SCIENCE – MASTERS PROGRAM

Contact info: goliveir@syr.edu - 302-727-1599