Jason “Scott” Person

Learning Response 4.3: Titanic Feature Engineering Strategies

20250429

# Overview

Review these assigned articles and notebooks, respond to the questions below, and develop a list of strategies to enhance our machine learning results with the Titanic data set.

# Required Learning Resources

1. **Machine Learning with Kaggle: Feature Engineering – Datacamp**<https://www.datacamp.com/community/tutorials/feature-engineering-kaggle>   
   ***NOTE:*** This article is an easier read between the two and a good warmup. It is also less detailed, leaves out a key step for brevity, and does not achieve a high accuracy rate.
2. **Gunes Evitan, Titanic - Advanced Feature Engineering Tutorial – Kaggle Best Submission 2020**  
   <https://www.kaggle.com/gunesevitan/titanic-advanced-feature-engineering-tutorial>   
   ***NOTES:*** 
   * Evitan’s notebook achieved a top score with the Titanic Dataset. In a competition that’s been going on a long time, the feature construction steps he demonstrates paid off. You might observe that his accuracy score of 83.7% does not seem that much higher than our results, but in predictive modeling. the difference between average outcomes and exceptional outcomes can often be just a few percentage points. And getting those last few percentage points can take a fair amount of work!
   * I don’t expect this much detail in this current course. ... However, do note the level of subject matter research and careful strategic thought Evitan has put into feature construction. This speaks to the key role of *domain knowledge* in data work. In a business environment, we make sure to bring a subject matter expert into the team. Even so, as McCormick discussed, doing additional research into the stories behind our data can make a big difference.
   * Evitan has null Fare values, where we did not. I believe this is because he combined the Kaggle train and test sets into one, whereas we have been working with the train set.
   * Evitan’s dummy (or one-hot) encoding process can be slightly improved by dropping one of the new columns. For instance, we only need one Sex column (0 or 1), and one fewer columns for the other categorical variables.
3. **Gunes Evitan, reply to Yoni Krichevsky in the comments beneath the above notebook.** This link should take you directly to Evitan’s reply. Topics include: rationale for using median for Fare and Age, how the number of bins for Fare and Age were chosen, etc.  
   <https://www.kaggle.com/gunesevitan/titanic-advanced-feature-engineering-tutorial#700295>

# Response Questions

1. Build a numbered list of all the feature engineering steps recommended in these resources. For each step, provide a short 2-7-word summary name and a one-sentence description.  
   * Exploratory Data Analysis (EDA) – Examine the data for shape, missing values, correlations etc. to discover the value of available data features and issues.
   * Data Cleansing – Deal with the issues discovered in EDA like missing values, duplicates, and inconsistencies.
   * Transform Data – Using encoding, scaling, and normalizing techniques to enhance the data set for the modeling algorithm.
   * Create New Features -
     + Bin Continuous Features – Combine continuous features into roughly equal sized groups (bins).
     + High Cardinality Grouping – Group high cardinality ordinal and nominal variables into groups like the binning of continuous features.
     + Extract Features – Using domain knowledge, extract information from smart keys and text columns.
   * Select Features – Remove features that are redundant or offer little or no value to the algorithms.
2. Evitan adopts a more targeted approach to filling values for Age. Summarize that approach.

Evitan calculated the median age for each Pclass (Passenger Class) and Sex group. This was then joined to records with missing values by these features. This process allowed for a more accurate filling of the missing values.

1. Evitan creates binned versions of Fare and Age, and a grouped version of Family\_Size. He does this for a purpose. What is that purpose? (Also, if you are not familiar with binning, look it up! It’s a common process and can offer significant advantages in many data-analytical contexts.)

For fare particularly Evitan got significant benefit from binning, which smoothed out the skewing and outliers on the high end. For all the fields, binning allows for a simpler model and a better representation of the nonlinear relationship between the predictor and target variables. All three variables have middle bins that had higher than expected survival rates.

1. Did the step of extracting titles from the Name field surprise you? There’s a lesson in feature construction we should draw from this. Reflect on this for a bit, and then (a) describe the kind of value you think may be derived from the new title field for training our machine learning model, and then (b) give a shot at drafting a statement of a lesson we should take away from this example and bear in mind as we work with other data sets.

*As you reflect, consider:*

* 1. Jedamski simply had us drop the Name field as irrelevant to prediction. Of course, Jedamski’s LinkedIn courses are beginner-level course. We are now moving into more advanced concepts and opportunities ...
  2. In McCormick’s terms, the Name field is a clear example of an ID field, typically not useful for predictive modeling. But remember that McCormick shared a few examples in which meaningful information can be extracted from ID fields ...
  3. Finally, recall how strongly McCormick stressed the importance of feature construction — the creation of new fields by strategically leveraging information available but underutilized in other fields.

The title field would likely correlate well with income. Those with “special” titles would tend to be of higher socioeconomic class and I would anticipate correlate with a better survival rate. I posit that those with more money and status commanded a higher status and thus preferential position when the decision was being made as to whom would be allowed onto the lifeboats. Title is also indicative of sex, which would obviously correlate with that feature as well as survival rate. Interestingly the model uses both. Title along with sex would give a better differentiation (more granular) than just sex.

The takeaway here is that text fields and other unstructured data can contain valuable features. Evitan proved that extracting the title and grouping similar values provides significant benefit to the model. As mentioned above in the instructions, small percentage gains in model performance are the difference between average and exceptional models.

### 