Analysis of Passenger Counts vs. Ticket Validations - Cooperation with Austrian Railways

Dávid Chmelík, Yuka Obayashi, Tomáš Tax

**Challenge**
1. **Main goal:** Estimate missing counts
   - Conductor counts are missing in 1/3 of trains
   - Train a machine-learning model that predicts counts
   - Use it to fill in the gaps in counts
   - Compare to existing model: Average counts on the segment

2. **Secondary goal:** Exploration of validations
   - Relationship with counts
   - Segment length, regional effects
   - Delays, different train categories

**Data Understanding & Data Preparation**

<table>
<thead>
<tr>
<th>Data Manipulation</th>
<th>Feature Induction</th>
<th>Data Cleaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aggregating individual validations into counts</td>
<td>Delays, segment length (in minutes)</td>
<td>Allow model validation and evaluation: keep only instances with available counts</td>
</tr>
<tr>
<td>Segment size (count)</td>
<td>Occupancy ratios (count/capacity)</td>
<td>Select features for modeling, avoid data leakage</td>
</tr>
<tr>
<td>Total validation count per train ride</td>
<td>Average count per ride (info from neighbor segments)</td>
<td>Many categories (segments): One-hot encoding</td>
</tr>
</tbody>
</table>

**Modeling**

**Baseline Model:**

\[
\text{BASE}_{i,t,x} = \frac{1}{n_{i,t,x}} \sum_{i=1}^{n_{i,t,x}} \text{COUNT}_{i,t,x} \]

- **AWS count for a weekday, train, segment**
- **Simulates current model used in production**

**Model Tuning:**

- **Model Tuning - 10 months of data**
- **Model Evaluation - last 2 months of data**
- **Regularization Regression - Lasso, Ridge, Elastic Net**
- **Ensemble Methods - Gradient Boosting, Random Forest, XGBoost**

**Workflow**

1. Set the project goal and the project plan. Understand the time between several datasets.
2. Merge and aggregate the data, introduce source features for the data exploration and modeling.
3. Conduct data exploration focusing on secondary goal of the project.
4. Cleanse the data for modeling; handling missing values in counts and selecting relevant features.
5. Setup the baseline model and define the evaluation framework for the modeling.
6. Implement multiple machine learning models, improving hyperparameter tuning.
7. Evaluate the performance of the models and provide interpretation of the results.
8. Select the currents for presentation and impact. Prepare the final deliverables.

**Result**

**2nd Class (SC):**

**Improvements:**

<table>
<thead>
<tr>
<th>Model</th>
<th>MAE</th>
<th>MRE RSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>48.9</td>
<td>1082.88%</td>
</tr>
<tr>
<td>Rijeka</td>
<td>39.9</td>
<td>763.72%</td>
</tr>
<tr>
<td>Ljubljana</td>
<td>39.0</td>
<td>763.73%</td>
</tr>
<tr>
<td>Linz</td>
<td>39.0</td>
<td>763.72%</td>
</tr>
<tr>
<td>Graz</td>
<td>39.0</td>
<td>763.72%</td>
</tr>
<tr>
<td>Trier</td>
<td>39.0</td>
<td>763.72%</td>
</tr>
<tr>
<td>Xpress</td>
<td>39.0</td>
<td>763.72%</td>
</tr>
<tr>
<td>GradientBoosting</td>
<td>37.0</td>
<td>750.71%</td>
</tr>
</tbody>
</table>

**Setting**

- **SC** Random Forest 29% 30%
- **SC RJ** Random Forest 29% 17%
- **RC RJ** Ridge Regression 1% 14%

**Conclusion**

In conclusion, we successfully achieved two goals of our project.

**Main goal:** Estimate missing counts
- We implemented multiple machine learning models, incorporating hyperparameter tuning and using cross-validation to identify optimal parameters for each model. And we finally developed the Random Forest Regression model, achieving a remarkable 30% increase in predictive accuracy over the baseline model.

**Secondary goal:** Exploration of validations
- We analyzed the number of validations with respect to certain factors.

**Further Research: Ideas to enhance the performance of our model**
1. Create a questionnaire for conductors to understand reasons behind missing counts.
2. Identify error values entered by conductors, considering entries exceeding capacity by more than 50%.
3. Improve the accuracy of the capacity data for trains with two components (double heading).
4. Include regional dummies in the model.