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

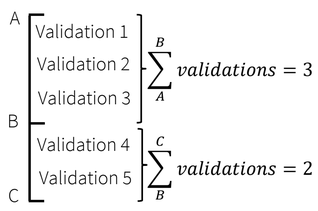
Data Understanding & Data Preparation

<u>Data Manipulation</u>

- Aggregating individual validations into counts
- Segment size unification for counts and validations

Original data format:

- - Count from A to C
- - Individual validations
- from A to B and B to C
- → Unify the format



<u>Feature Induction</u>

- Delays, segment length (in minutes)
- Occupancy ratios (count/capacity)
- Average count per ride (info from neighbor segments)
- Total validation count per train ride

<u>Data Cleaning</u>

- Allow model validation and evaluation: Keep only instances with available counts
- Select features for modeling, avoid data leakage
- Many categories (segments): One-hot encoding

Modeling

<u>Baseline Model:</u>

$$BASE_{w,t,s} = \frac{1}{n_{w,t,s}} \sum_{i=1}^{n_{w,t,s}} COUNT_{w,t,s,i}$$

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

<u>4 Model Settings:</u>

2nd Class (SC) | 1st Class (FC) | 2nd Class (SC RJ) | 1st Class (FC RJ)

<u>Pipeline:</u>

ML Model → Model Tuning (MAE↓) → Evaluation

- 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

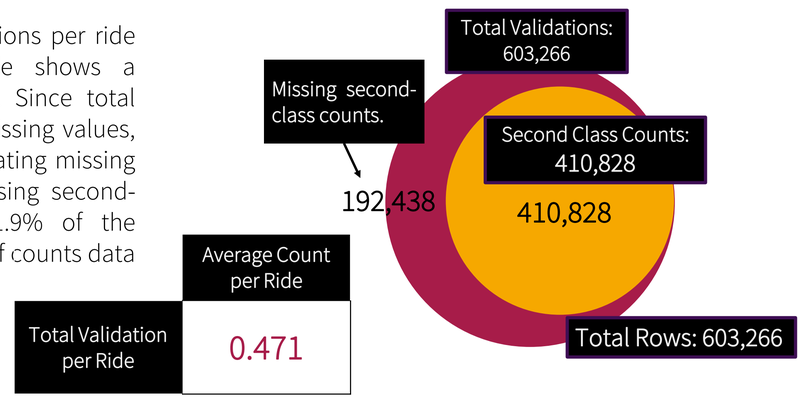
Dataset

- **Validations** – Information about individual ticket validations together with the ticket number, train number, and date of validation.
- **Tsdata** – Contains all tickets sold through the ÖBB ticket shop. Allows to retrieve comfort class of validated tickets.
- **Counts** – Research target. Includes conductor counts for the first and second classes, a train number, date, and segment codes.
- **CountSegments** – Defines the count segments. Crucial for comparing validations and counts (different segment granularity).
- **Capacities** – Contains information about what seating capacity a train has in first and second class on a given date.
- **Fahrplan** – Schedule of all trains. Includes information about the times of arrival and departure of trains to their stations. Moreover, this dataset includes interesting insight in the form of the train category, for example, RJ, RJX, IC, ICE, etc.
- **Istfahrten** – Source for calculating delays as it contains scheduled and real times of departure of a train from a station.
- **Stations** – Details (name of the station, district, state, and its longitude and latitude) about a station from its code.

Data Understanding & Data Preparation

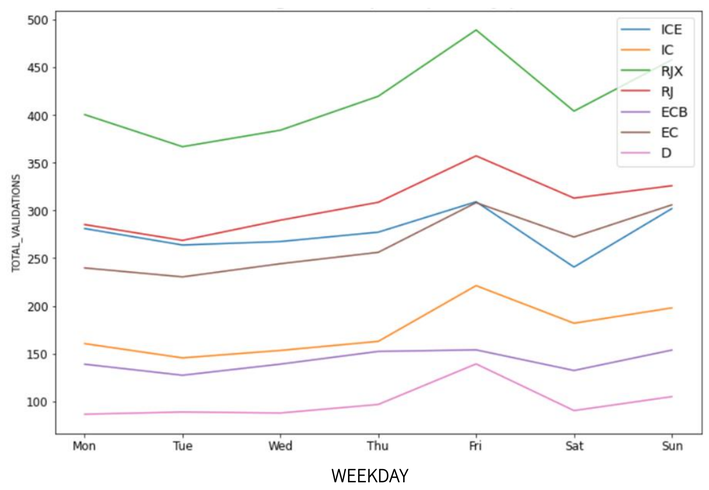
<u>Estimate missing counts</u>

Our new features, total validations per ride and average count per ride shows a moderate positive correlation. Since total validations do not have any missing values, they can be valuable for estimating missing values in the count data. Missing second-class counts account for 31.9% of the dataset. In the first-class, 35% of counts data is missing.

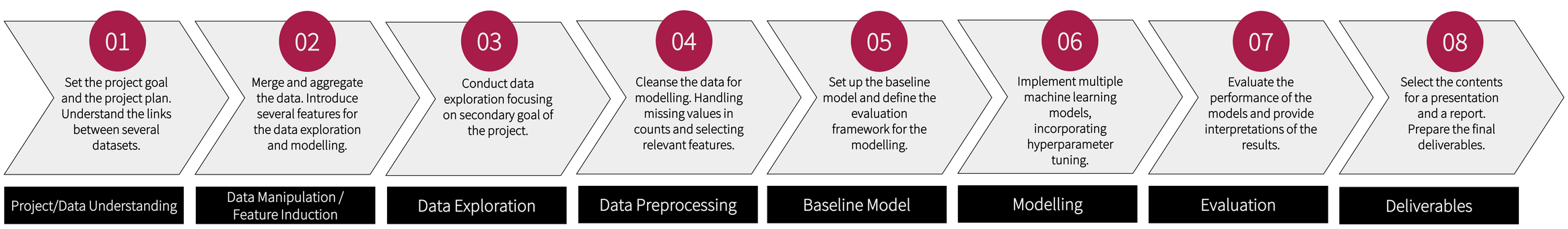


<u>Exploration of validations</u>

Temporal Effects/Train Category: There is increased traffic, particularly around the weekends, with peaks on Friday and Sunday, and a dip on Tuesday and Wednesday. Railjet Xpress (RJX) has the highest number of total validations while D is the lowest.



Workflow



Result

2nd Class (SC):

Model	MAE	MSE	R2
Baseline	48.9	10828.8	38%
Lasso	39.9	7663.7	56%
Ridge	38.3	7433.2	58%
ElasticNet	39.0	7530.0	57%
RandomForest	34.9	5591.7	68%
GradientBoosting	35.1	5370.1	69%
XGBoost	36.6	8228.7	53%

Improvements:

Setting	Best Model	MAE	R2
SC	Random Forest	29 %	30 %
FC	Random Forest	6 %	30 %
SC RJ	Random Forest	29 %	17 %
FC RJ	Ridge Regression	1 %	14 %

Conclusion

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

<u>Main goal: Estimate missing counts</u>

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.

<u>Secondary goal: Exploration of validations</u>

We analyzed the number of validations with respect to certain factors.

<u>Further Research: Ideas to enhance the performance of our model</u>

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.