Capstone 2 Project Report

1. Problem Statement

This data science project aims to create a predictive model for whether a shift within a Cleveland healthcare providers' system will be canceled or not, using shift characteristics such as shift start times, shift length, type of healthcare worker, compensation, and facility.

The healthcare provider operates 67 facilities that employ or contract four main types of healthcare workers: certified nursing assistants, licensed vocational nurses, nurses, and registered nurses. The shift assignment for these workers is done partially through the software of a staffing agency that allows medical personnel to book or cancel shifts with any of the facilities. From October 2021 through January 2022, the high number of shifts canceled on short notice potentially posed problems for the healthcare provider such as difficulty to call in backup, high workload for remaining employees, and impact on patient quality of care.

The challenge is to predict shifts that may be canceled (binary classification) with 80% accuracy or better, allowing the healthcare provider to make contingency plans. Another goal is to identify the top three parameters that drive shift cancellations for the healthcare provider to further investigate and address.

2. Data

Three CSV files with shift data from the shift booking software are available on Kaggle: for shifts, bookings, and cancellations. Shift data include shift ID, facility ID, shift start and end date and time, shift type of am, pm or night, type of healthcare worker requested, the hourly charge, and the number of paid hours (shift time). Booking data include shift ID, facility ID, worker ID, record creation date and time and lead time to shift start. Cancellation data include the same fields as bookings, plus a field indicating whether it was a ahead of time cancellation or a no show.

Kaggle data, GitHub data

3. Data Cleaning & Wrangling

The time window of shift dates that appeared in all three data sets (shifts, bookings, and cancellations) was 10/01/2021 to 11/30/2021, so any rows outside this window were removed. This reduced the number of facilities included in the analysis to 49. The shift table was used as the 'ground truth' for bona fide facility IDs, as there were over 900 in both bookings and cancellations (possibly through data entry error or software problem).

Since the raw shift data only included basic information about shifts, such as start time, end time, hourly pay rate, healthcare worker type and facility ID, some additional derived features were created, such as total compensation (taking into account a cut of 22% for an entity identified as CBH - possibly the staffing agency), day of week and holiday indicator for 6 days around Thanksgiving, and hour of day the shift started.

For bookings and cancellations, day of week and hour of day features were added for the time the booking or cancellation was made and also for the shift time itself. A holiday indicator was added to both data sets. Start date and time of the booked shifts was calculated from time the booking was created and lead time. Also, a column containing just the date of the shift was added to both bookings and cancellations. To cancellations , a 'under 24 hour lead time' indicator was added.

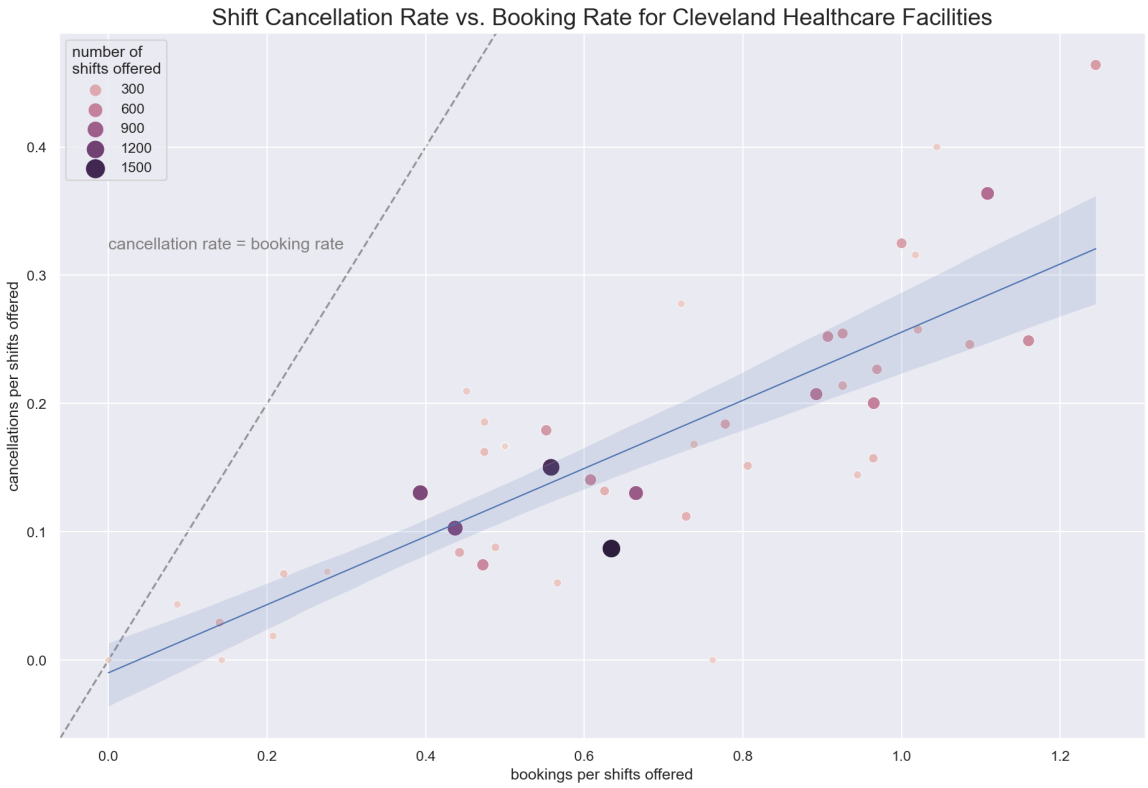
A total of 147 rows were removed because of inconsistent date and time fields. The only rows with missing values were not dropped, since the column containing NaN values (worker ID) was removed. While all numerical features had outliers, none could be identified as obvious mistakes. The target feature 'canceled' was added to both bookings and cancellations, and set to 1 or 0, respectively. After retaining only common columns, the two tables were combined. Next, all but the most recent rows for shifts that were booked or canceled multiple times were removed, and the result was stored in a CSV file, containing 9176 rows and 28 columns.

A summary table for the 49 facilities was created containing the number of shifts, bookings and cancellations for each facility (4 had no bookings, 6 had no cancellations, filled with zero). Also, cancellation rates (number of canceled shifts / shifts) and booking rates were calculated.

clean & wrangle & EDA notebook, cleaned data (csv)

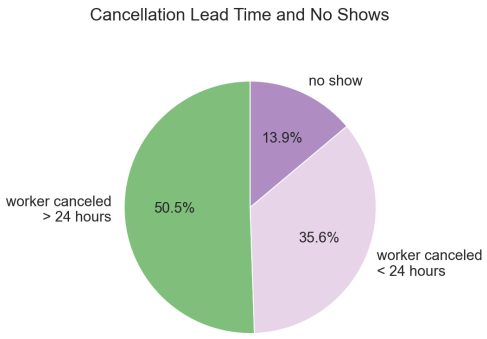
4. Exploratory Data Analysis (EDA)

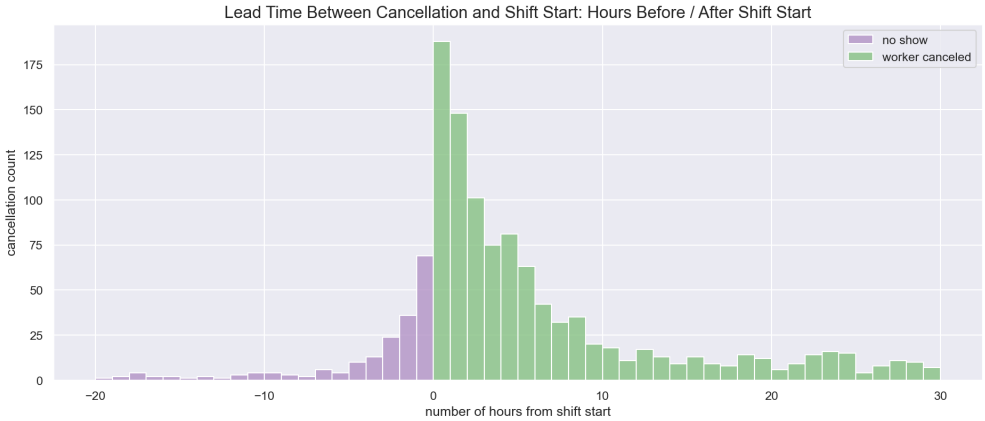
A heatmap of features showed some (but not perfect) expected correlation between total compensation and hourly charge and shift length. Pairplots of these numerical features using select categorical variables as hue indicated one derived feature as highly correlated with shift length (short shift indicator, not mentioned in 3. since it was subsequently removed).



Based on the facilities summary table, a plot of facilities' cancellation rate vs. booking rate with linear regression and 95% confidence interval shows that facilities scatter fairly widely around the line. The few larger facilities (going by number of shifts offered) seem to fall into the 'low-to-medium churn' region, and all facilities have higher booking rates than cancellation rates. A similar plot for worker types shows that for nurses (not registered nurses) the cancellation rate is as high as the booking rate, and that nursing assistants are the group with the highest churn (high booking and cancellation rate). Summary statistics comparing shift length and hourly compensation between shifts offered and shifts canceled show the latter have slightly higher mean length and compensation, at least for nursing assistants and vocational nurses.

Analysis of lead time of cancellations shows that the portion of no-show cancellations is considerable: about 14%. Additionally, cancellations that happen with less than 24 hours lead time make up almost half of all cancellations (the 14% no shows are included here). The number of cancellations increases the closer it gets to the time the shifts start.





[clean & wrangle & EDA notebook]

5. Pre-processing and Training - Model Selection

Excluding administrative columns or features with no predictive value or high cardinality, such as worker IDs and lead time, three numerical features (shift length, hourly charge, total compensation) and six categorical features (facility ID, worker type, day of week, hour of day, shift type, holiday shift indicator) were retained for modeling. Data was split into training and test sets with a 80-20 ratio. The positive (canceled) class to negative class ratio was 1:6, so gridsearch for various classification models was performed using stratified cross validation with five splits. The pipeline included a robust scaler for numerical features and one-hot encoder for categorical features, and the scoring method was ROC-AUC.

Model	Train Score ROC-AUC	Test Score ROC-AUC	Parameters	Feature Importance
Decision Tree	0.74	0.75	criterion: gini; max_depth: 7; max_features: 0.9; max_leaf_nodes: 15	Shift length Workter type nurse Facility ID ba39 Facility ID f82c
Decision Tree (unscaled)	0.74	0.75	criterion: gini; max_depth: 7; max_features: 0.85; max_leaf_nodes: 15	Shift length Worker type nurse Facility ID ba39 Facility ID cbcd
Random Forest	0.79	0.80	max_depth:10; n_estimators: 200	Shift length Total compensation Worker type nurse Hourly charge
Gradient Boosting (CatBoost)	0.84	0.84	depth: 4; iterations: 300; l2_leaf_reg: 5; learning_rate: 0.1	Shift length Total compensation Worker type nurse Hourly charge
Logistic Regression	0.70	0.72	C: 1.0; max_iter: 100; penalty: l2; solver: newton-cg	Worker type nurse Night shift Facility ID ending ebff Facility ID ending 7b4b

All models show slight overfitting. For Decision Trees, the dominating feature was shift length, followed by major features, such as worker type nurse, and two particular facility IDs. Without scaling, the only change is in the importance of the minor features. The resulting tree shows two distinct **cutoff shift lengths** that determine the split between the booked and canceled class: At about **3.5 hours and 7.5 hours, longer shifts lead to**

more cancellations. Both tree ensemble methods showed shift length, total compensation, nurse, and hourly charge as the important features. Interestingly, Logistic Regression shows worker type nurse as the most important feature (positive sign, favoring cancellation) and shift type night as a second (negative sign, favoring booking), with shift length not appearing at all. However, it was the weakest model in this comparison. Between all models, a common important feature was **worker type nurse**. According to the Decision Tree models, this worker type **leads to more cancellations.** Test metrics (see table) and additionally collected accuracy and F1 scores point to CatBoost as the best model.

pre-processing & modeling notebook, model metrics (csv)

6. Modeling

The winning model was CatBoost Gradient Boosting, which was subsequently run in 'native' mode, that is, without prior scaling and one-hot encoding. This resulted in a slightly better performance, with ROC-AUC scores of 0.85 and 0.86 for train and test, respectively. Compared to the pre-scaled and one-hot encoded version, the only changed model parameter was l2_leaf_reg: 1. Its **most important features, in order, were shift length** `(score 57.6), total compensation (score 13.3) facility ID (score 10.7), and worker type (score 6.5).` The model's **accuracy on test data was 73%**

pre-processing & modeling notebook

7. Recommendations

Since the consistently most important feature was shift length, and because there seem to be clear thresholds around 3.5 and 7.5 hours beyond which cancellations occur, this should be an important factor to consider in shift planning and contingency planning. The underlying reasons for these cancellations may be workers' difficulties to attend to family responsibilities or other obligations, such as a second shift with another employer, when they take on shifts that last longer than a typical half day or full time work day. The healthcare provider could **prioritize shift length below the thresholds of 3.5 and 7.5 hours** and **add 'on call' contingency personnel for non-standard shift lengths**.

Additionally, nurses (not licensed vocational nurses or registered nurses) over-proportionally cancel shifts. Comparing the box-plot of their hourly charges with that of the other worker types, they are at the low end of the range for licensed vocational nurses, and just above the range for certified nursing assistants. Their compensation may reflect the healthcare providers need to fill positions that do not require licensure, even though in practice these shifts are manned by nurses that may have these certifications. They may cancel as soon as they find a better paying shift, so it could be worth **investigating whether a higher hourly charge for nurses on shifts that do not require licenses** could prevent this.

8. Future Work and Limitations

The data contained some administrative fields with unclear purpose but potential implications for the validity of the data, for example, shift data contained a field indicating whether a shift was deleted (presumably by the facility), sometimes despite the same shift appearing in the booked shifts table. An effort was made to remove as many of these inconsistencies as possible. The Logistic Regression model may improve with centering the data, which was omitted in favor of pipeline re-use. Since some models point to particular facilities leading to more cancellations than others, these could be identified and investigated further. Additional data, such as type of facility, e.g., hospital vs. nursing home, would be helpful for this. Also, data collection cutoff time appears to be determined by data a shift, cancellation, or booking was entered into the table, instead of using the date of the shift. This leads to misalignment that cut the available time frame for analysis from four to just two months.

9. Conclusion

The accuracy achieved was only 73% compared to the goal of 80%. However, some useful information as to which factors contribute to cancellations could be extracted, such as shift length, worker type, and facility ID. The latter should be analyzed further, ideally with more data collected specifically for facilities, such as type and actual size.