Section 1: Introduction/Summary

The data I'm using is Citywide Payroll Data obtained from NYC OpenData. It is provided by the Office of Payroll Administration and was last updated on October 30, 2024. Each observation in the dataset is an employee of the NYC municipal government and information is recorded about their agency, borough, and job title as well as compensation information, the amount of overtime they worked, etc. Note that the compensation data for employees (i.e. final base salary and gross salary) are recorded at the end of each fiscal year—this aligns with my use case. To be clear, my report makes it easier for the heads/managers of large agencies at the NYC municipal government to sit down at the end of the fiscal year and anticipate who is going to leave their agency, informing their staffing plans for the year ahead.

Thus, the research question I'm aiming to answer is **can we predict whether a civil-servant will stay on the payroll next year (t+1) with reasonable accuracy (greater than 80%), given information we have about their seniority, agency, salary, and other engineered features at the end of this year (t)**. Because I am theoretically giving this report to agency heads as a supplementary piece of information for them to use, I would rather over-estimate than underestimate churn, so I prioritize recall when tuning my model.

First, I create a definition of employee churn and narrow my investigation to 3 of the largest agencies (that don't primarily house temporary workers): the Department of Parks and Recreation, Department of Sanitation, and Department of Correction. An employee is defined as churned when their first name, last name, and middle initial disappear from the payroll of their agency, borough combination in the next year. So individuals that are promoted or change titles within agencies are rightfully *not* counted as churned. However, individuals' whose entries are discontinuous on the payroll (i.e. they leave an agency and then come back) are counted as churned and registered as a new employee when they return. The logic behind this being that as an agency head, you have to replace the individual for that role while they are gone, so predicting when they leave (even if to eventually return) is valuable information.

As I've said, the intended audience for my research are the agency managers of the aforementioned departments who need to decide A) approximately how many new hires they will have to account for time and budget wise and B) who to give extra bonuses to in order to incentivize their retention. I use a **Random Forest model** to predict churn, using the following metrics in evaluation: Accuracy (total % of employees predicted correctly), Precision, and Recall (evaluates how good the model is at actually identifying all churned employees). I intended to compare the RF model to a baseline survival model (cox-proportional hazard) to determine if the machine learning model does better than a dumb and reasonable classifier. I used seniority as a proxy for time since hire and churn as the event indicator. However, the survival model was not behaving as expected (it simply predicted everyone as churned) because I didn't censor observations that weren't churned by 2024 (last year in the dataset). In the future, I would fix this issue and compare against the Random Forest model as planned. But in this report, **I compare against a baseline model that randomly guesses churn using knowledge of the underlying proportion of retained:churned employees in each training set.**

I engineer the following features for my analysis: hourly_change_pos_log, ot_change_signed_log, ot_pay_quintile, and ot_vs_ma3_prev (overtime hours this year divided by moving average of overtime in past 3 years; I did not end up using this particular feature in my modeling because it was not able to be calculated for around 50% of the observations in the dataset). hour_change_pos_log calculates each employee's base_salary in terms of hourly rate, designating if the amount increased from last year and the difference it increased by. Employees that experience a decrease in hourly rate, no change in hourly rate, or are new to the payroll have value recorded as 0 for this feature. ot_change_signed_log determines the change in an employee's overtime hours compared to last year. For this feature, negative values remain as they are and aren't coerced to 0. Employees new to the payroll still have value recorded as 0. ot_pay_quintile takes an employee's total overtime pay and determines what quintile of the overall overtime pay distribution they're in. It does not calculate agency specific distributions, which would have been ideal with more time. While I will provide additional explanation for why I made specific decisions regarding these features below, my overarching goal was to exploit information regarding an employee's incentive to stay (any increase in pay in the last year), their commitment to an agency via comparison to historical behavior (overtime hours this year vs last year), and their commitment to an agency via comparison to peers (where they fall on the overall, overtime pay distribution).

The key finding I obtained from my final Random Forest model that uses 2015-2023 data in its training set, and predicts on feature data from 2024 suggests the following estimates for employee churn in 2024:

A. **Overall percent to churn in 2024 across the 3 major agencies : 30.2%**
B. **Department of Parks and Recreation: 64.3%**
C. **Department of Sanitation: 10.6%**
D. **Department of Correction: 6.0%**

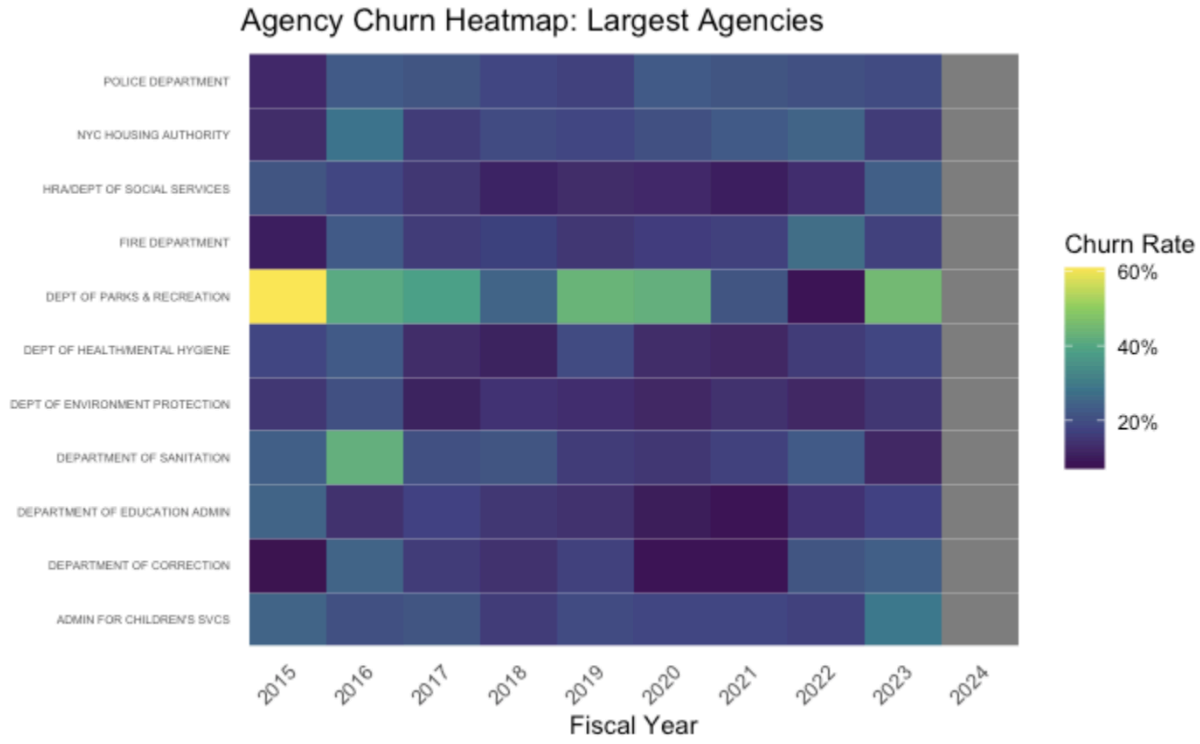
I will discuss the uncertainty associated with my model later on in the report.

Section 2: Data Cleaning, Exploration, and Justification for Chosen Agencies

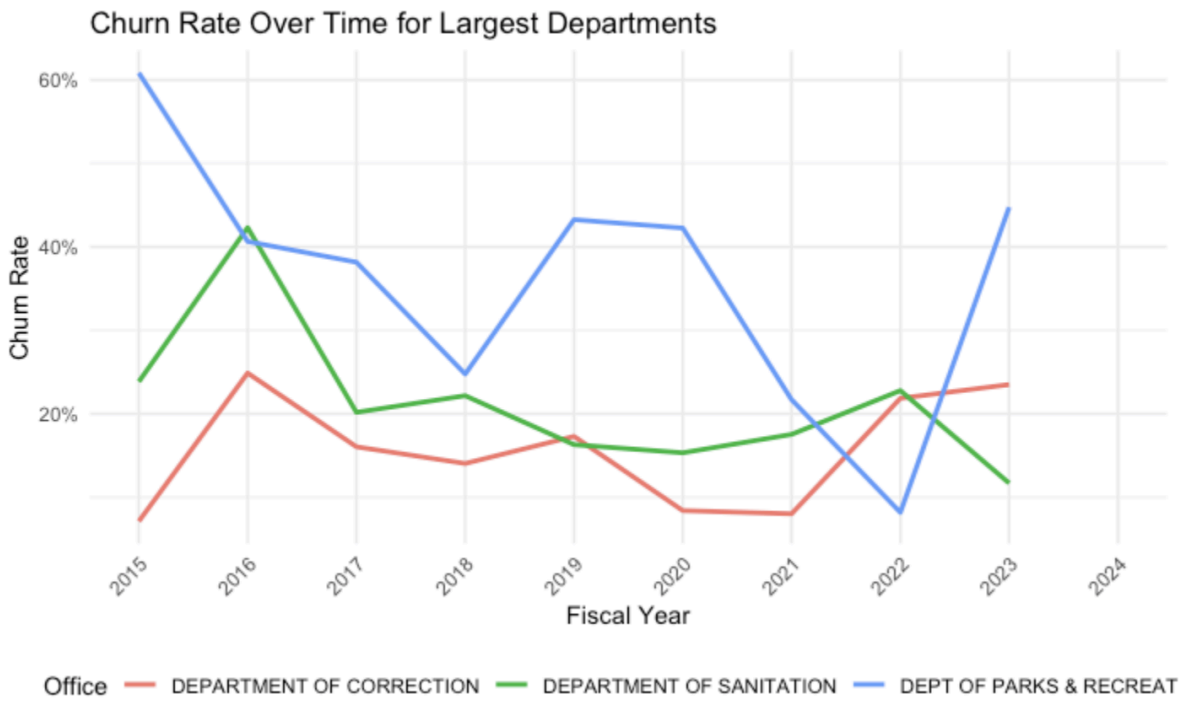
To track all unique employees in the dataset, I created a key that consists of each observation's first name, last name, middle initial, agency name, and the borough they work in. This variable is designated emp_key. All observations that did not contain a value for first_name or last_name were dropped. Then, I checked to see if there were cases of the same emp_key occurring multiple times in one year. Some instances of duplicate emp_keys appear to be the result of multiple people having the same name in an agency, borough. Through visual inspection, I decided that if there are more than 5 duplicates in one year that likely implies that multiple people have the same name in an agency, borough and removed those observations (only 535 observations were removed from this condition). Next, the observations that had 2-4 duplicates in a given year appeared mostly to be the result of people who held multiple titles concurrently under one agency (i.e. an adjunct professor who also serves as IT support) (see Appendix 1). Accordingly, I assumed these entries to all belong to the same person, created 2

new columns: total_rows and n_titles so I could flag them, and collapsed the multiple entries by only keeping one for each emp_key, year combination. This way, these observations will not cause unnecessary noise to enter my churn calculations. Notably, duplicate emp_keys in a fiscal_year constituted a very small minority of total observations. **After this cleaning, I verified that each observation was now a unique emp_key, fiscal_year combination and computed churn_flag to = 0 if the employee is retained (shows up in next year's payroll), and = 1 if the employee is churned (does not show up in next year's payroll).**

I discovered a change in the coding of boroughs and select agencies from 2014 to 2015, which made comparisons prior to 2015 difficult. As a result, I removed observations from 2014 and only kept the years 2015-2024. Additional cleaning procedures are detailed in my code. Below is a visualization that shows the employee churn variation across years in the *largest* agencies (by number of unique employees). Light colors signify high churn, while dark colors signify low churn.



Of the largest agencies, the Department of Parks and Recreation, Department of Sanitation, and Department of Correction have the highest variation in churn across years (see Appendix 2). I visualize the time series for these departments below.



As the graph shows, churn for these departments, especially Parks and Recreation and Sanitation, is not only highly variable but can reach rates of 40-60% in certain years. Evidently, these agencies not only employ a large share of the civil servant workforce in New York, but they also demonstrate unstable retention patterns. My decision to examine these in the core dataset for my modeling is motivated by one main reason: **being able to reliably predict which employees will churn in 2024 for three large agencies dealing with unstable retention will generate the most value for my user compared to focusing on smaller agencies that can either handle the problem without machine learning methods or naturally have stable retention**. There are 3 supplementary reasons these are good agencies to choose: 1) Big agencies with large n increase the overall sample size compared to smaller agencies, making my model more stable 2) Big agencies also likely have more heterogeneity across predictors (which I explored in my code) 3) The years where these agencies have high churn help my model better understand what features are associated with positive cases. In combination, these 3 aspects improve the generalizability of my model. Finally, I removed roles that I knew were seasonal by: 1) Calculating average churn rate across years by title_description and removing titles that had 100% churn 2) Removing job titles that contained the terms “intern”, “helper”, or “trainee”.

I do note that because there are no small agencies in my sample, the model may not translate well to a small agency or specialized agency like the District Attorney’s office where the majority of people have salaried, desk jobs. Including such smaller agencies constitutes an alternative approach. But because I envision the use case of anticipating churn specifically being helpful for larger agencies with relatively variable Y-to-Y churn rates, I chose my given strategy. In my final dataset for the 3 agencies, there are 124,204 unique employee keys across the years 2015-2024.

Section 3: Feature Engineering, Justification for Chosen Model, and Cross Validation

The distributions of the categorical variables I include as predictors in my final model are explored in my code. **These categorical variables are: agency_name, work_location_borough, pay_basis, and other_pay_negative.** other_pay_negative is a binary variable that is 1 when an employee's total_other_pay is a negative value. Of more importance is the choices I made regarding the engineering of the remaining features in my model: seniority, hourly_change_pos_log, ot_change_signed_log, and ot_pay_quintile. Seniority was calculated to be the difference between the current fiscal year and an employee's agency start year. I already discussed how hourly_change_pos_log, ot_change_signed_log, and ot_pay_quintile were calculated in the introduction, but I want to provide further context regarding certain choices. 1) When calculating the change in an employee's *hourly* base salary for this year compared to last year, there were a few negative outlier values (i.e. min was -$71.5/hour) that I coerced to 0 to avoid unnecessary noise. The resulting distribution of the variable was very right skewed (Appendix 3), thus, I performed a log transformation. 2) When calculating the change in an employee's overtime hours this year compared to last year, the distribution of the raw change was more symmetric around 0, although still heavily spiked at 0 (Appendix 4). Because negative values appeared to be more indicative of actual changes in employee behavior rather than just a few random outliers, I decided to keep negative values as they are but still performed a signed log transformation to pull in the left and right tails. The final distributions for the hourly_change_pos_log and ot_change_signed_log features after log transformation are shown in Appendix 5 and 6.

I ultimately landed on using Random Forest primarily because a tree model automatically searches for non-linearities and interactions. It also deals well with the abnormal distributions in my engineered numerical predictors. Building a single tree means that predictions are very unstable. In my method of averaging 50+ trees, I am able to shrink that variability such that recall and precision estimates are more stable. I combine my Random Forest model with a time-based nested Cross-Validation procedure with 5 outer folds and 4 inner folds. That is to say, 5 unique splits are made to my data (using only the years 2015-2023). The first outer split tunes for hyperparameters on years 2015-2018 and tests the final model on year 2019. The second split tunes for hyperparameters on years 2015-2019 and tests the final model on year 2020, and so on. **My aim with this was to create a realistic forecasting setup that helps the managers of the 3 agencies understand how robust the model is to real drifts that occur in employee churn dynamics year-to-year.**

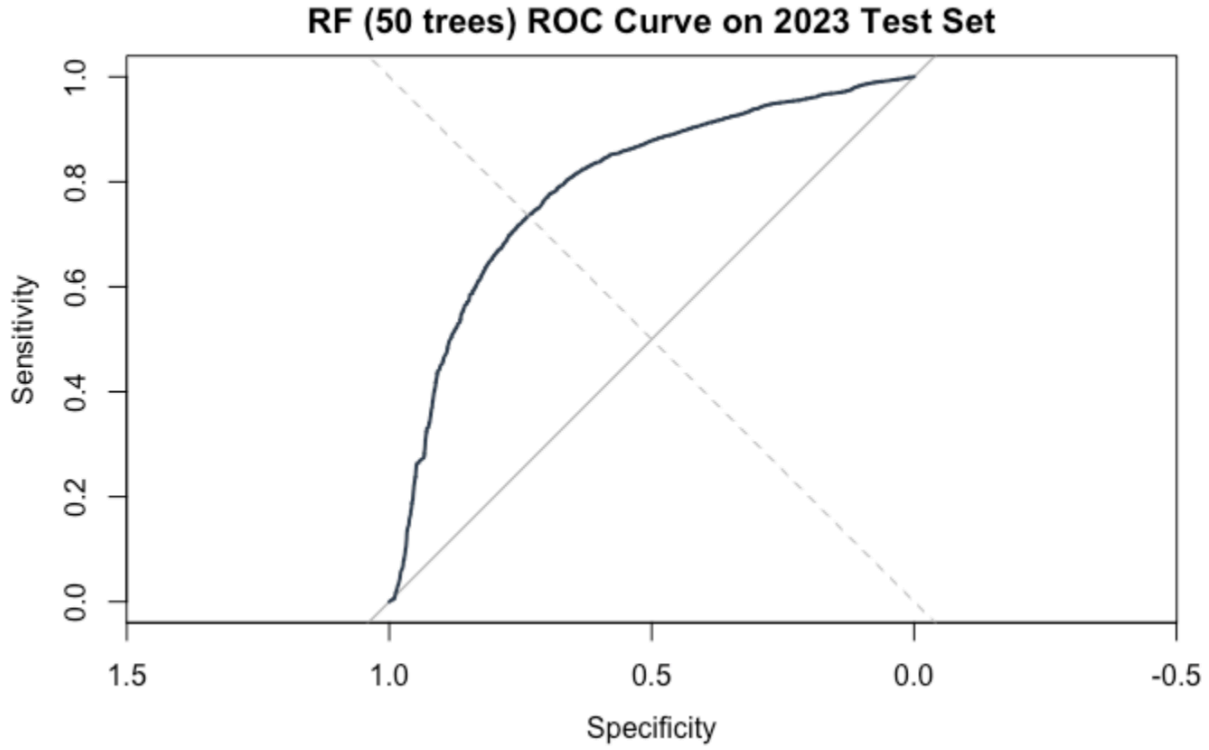
Section 4: Model Evaluation and Conclusion (Addressed to Agency Managers)

As shown by the barplots below, the Random Forest model does better than the baseline classifier (random guessing) in every test year 2019-2023. Crucially, the Random Forest model maintains recall near 70% every year, and it does so without sacrificing overall accuracy which remains at or above 70% year to year (Appendix 7). The recall rate tells us that the Random

Forest model is able to consistently identify true positive cases around 70% of the time, and that this is not the result of blatant over-prediction (or else we would observe much lower overall accuracy). These barplots show the results of the cross-validation procedure described above that excludes data from 2024; for this procedure, the threshold for what counts as a positive/churned case was set at probability > 0.3. However, as I will discuss below, you are free to further tune this yourself by visualizing the ROC curve. One thing to note is that in specific years—2021 and 2022—precision dips noticeably. This is likely because the churn for those years decreases across the 3 agencies in an unexpected way and the model is not able to fully understand why with the provided features. Overall, our Random Forest does a good job of generally predicting who's likely to leave—it gives a solid estimate of expected turnover to plan for next year. The accuracy, precision, and recall numbers display some variability year to year, but that is just an indication that there are some things it can't see like city-wide policy changes or economic shocks. To improve the year-to-year variability, we could grow more trees within each inner loop or bring in extra data to increase the features (context) our model has to train on.

Below is the ROC curve that you can use to adjust the threshold for what constitutes a positive/churned case. The ROC curve is the result of using the final tuned version of our Random Forest model, trained on 2015-2023 data, used to predict 2024 churn rate using 2024

predictor values. The more the curve bulges up toward the upper left corner the better our model's predictive power. The curve also visualizes the tradeoff between precision and recall, because even though we would rather overpredict than underpredict, balance is still necessary in the context of our problem. My suggestion using the ROC curve depicted below is to set the threshold equal to 0.256.



Finally, below is my official prediction for 2024's overall churn rate across the 3 agencies and churn rate by agency using the final tuned model. Refer to my Evaluation.Rmd code for the dataset containing predictions for each individual employee. Feature importance is broken down in Appendix 8 (not essential to my central focus of giving Correction, Sanitation, and Parks and Recreation managers a workable model for employee churn). Once again, additional data exploration, distributions for all features, model parameters/tuning information, and additional evaluation of results can be found in my code.

agency_name <fctr>	pct_churn <dbl>
DEPARTMENT OF CORRECTION	5.976806
DEPARTMENT OF SANITATION	10.631395
DEPT OF PARKS & RECREATION	64.344138

Appendix

Appendix 1

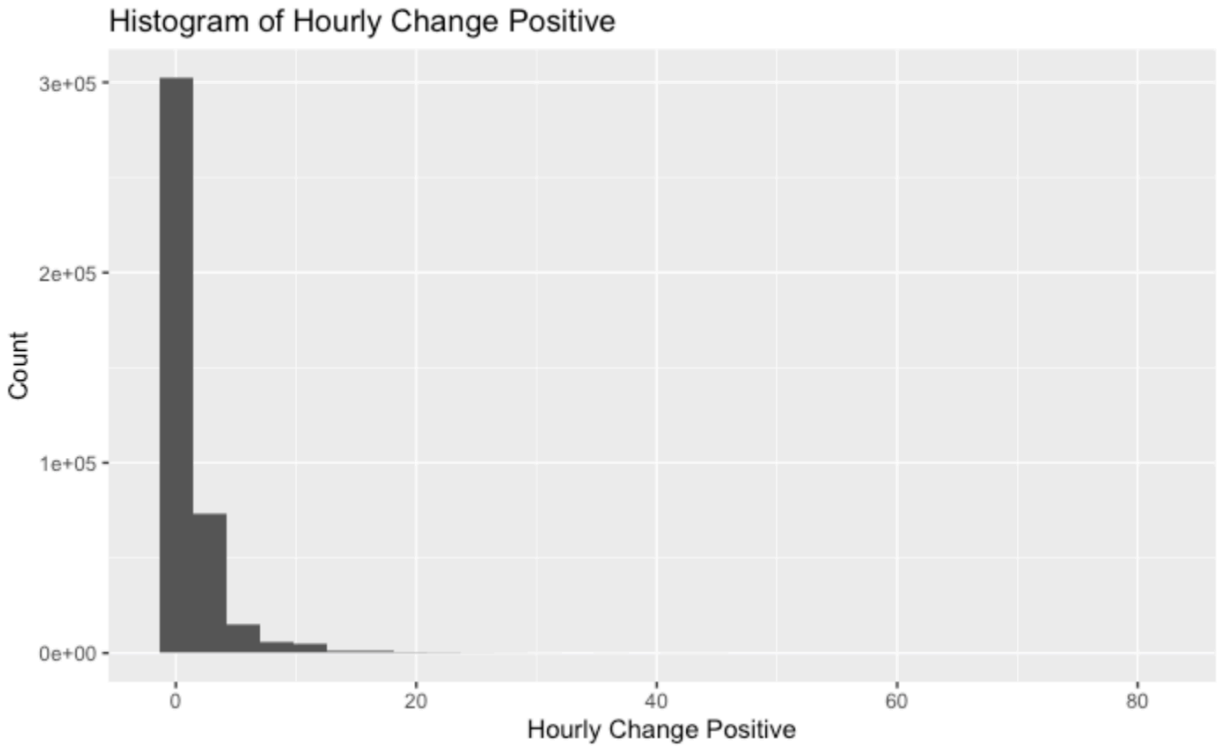
	emp_key	fiscal_year	job_titles	total_rows	n_titles
	All	All	All	[...]	All
1	ABDELLAH||AIT EL MOUDEN|COMMUNITY COLLEGE (L...	2015	ADJUNCT COLLEGE LAB TECH | ADJUNCT LECTURER | ...	4	4
2	ABEL|E|NAVARRO|COMMUNITY COLLEGE (MANHATTA...	2016	ADJUNCT ASSISTANT PROFESSOR | ADJUNCT LECTURE...	4	4
3	ABEL|E|NAVARRO|COMMUNITY COLLEGE (MANHATTA...	2018	COLLEGE ASSISTANT | NON-TEACHING ADJUNCT III | ...	4	4
4	ABEL|E|NAVARRO|COMMUNITY COLLEGE (MANHATTA...	2020	ADJUNCT LECTURER | NON-TEACHING ADJUNCT III | ...	4	4
5	ACHRAF|A|SEYAM|COMMUNITY COLLEGE (MANHATTA...	2019	ASSOCIATE PROFESSOR | CONTINUING EDUCATION TE...	4	4
6	ADELE||DOYLE|COMMUNITY COLLEGE (KINGSBORO)|B...	2019	CONTINUING EDUCATION TEACHER | NON-TEACHING...	4	4
7	ADOLFO||DEJESUS|COMMUNITY COLLEGE (BRONX)|BR...	2015	CUNY CUSTODIAL ASSISTANT | ADJUNCT LECTURER | ...	4	4
8	ADRIANA|L|BRADSHAW|COMMUNITY COLLEGE (LAGU...	2020	ADJUNCT LECTURER | COLLEGE ASSISTANT | NON-TE...	4	4
9	AISHAH||DEAN|COMMUNITY COLLEGE (BRONX)|	2014	ADJUNCT COLLEGE LAB TECH | ADJUNCT LECTURER | ...	4	4
10	ALEXANDER|R|MARTINEZ|COMMUNITY COLLEGE (QUE...	2017	COLLEGE ASSISTANT | IT SUPPORT ASSISTANT | ADJU...	4	4

Appendix 2

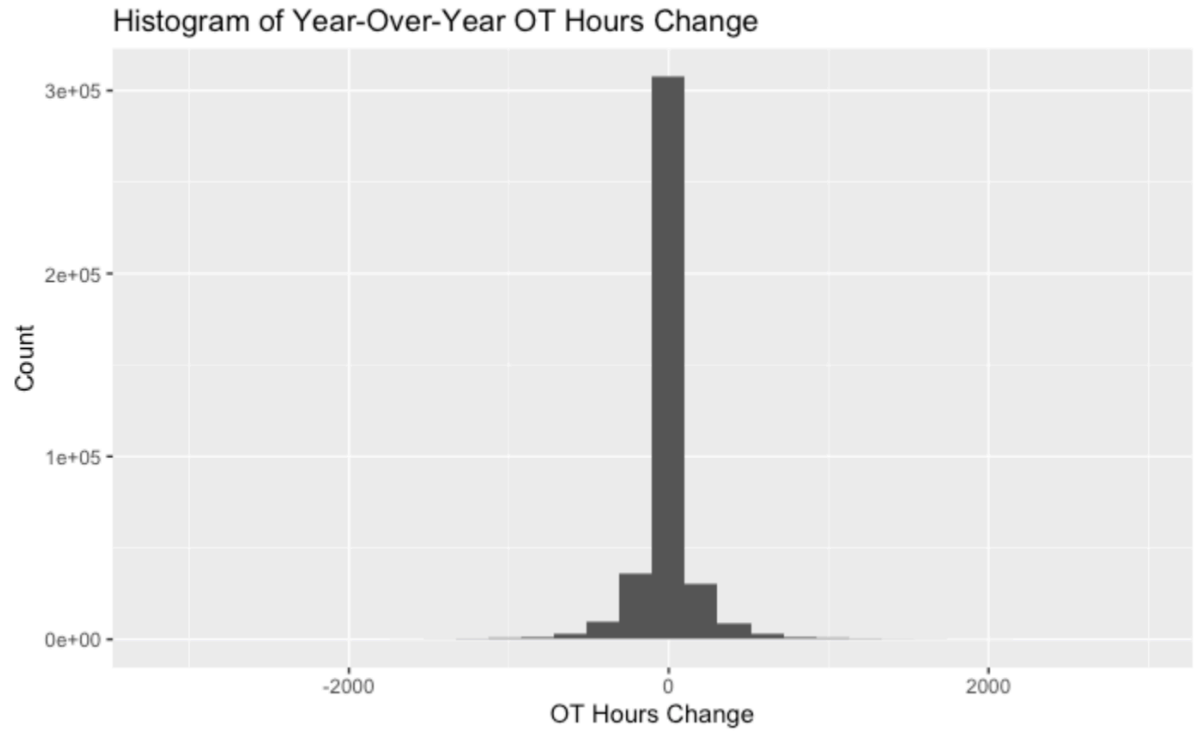
agency_name <chr>	var_churn <dbl>
DEPT OF PARKS & RECREATION	0.023944667
DEPARTMENT OF SANITATION	0.007734179
DEPARTMENT OF CORRECTION	0.004674891
NYC HOUSING AUTHORITY	0.002362899
DEPARTMENT OF EDUCATION ADMIN	0.002262718
FIRE DEPARTMENT	0.002232299
HRA/DEPT OF SOCIAL SERVICES	0.002172284
ADMIN FOR CHILDREN'S SVCS	0.001766688
DEPT OF HEALTH/MENTAL HYGIENE	0.001547982
POLICE DEPARTMENT	0.001194668
DEPT OF ENVIRONMENT PROTECTION	0.000715823

11 rows

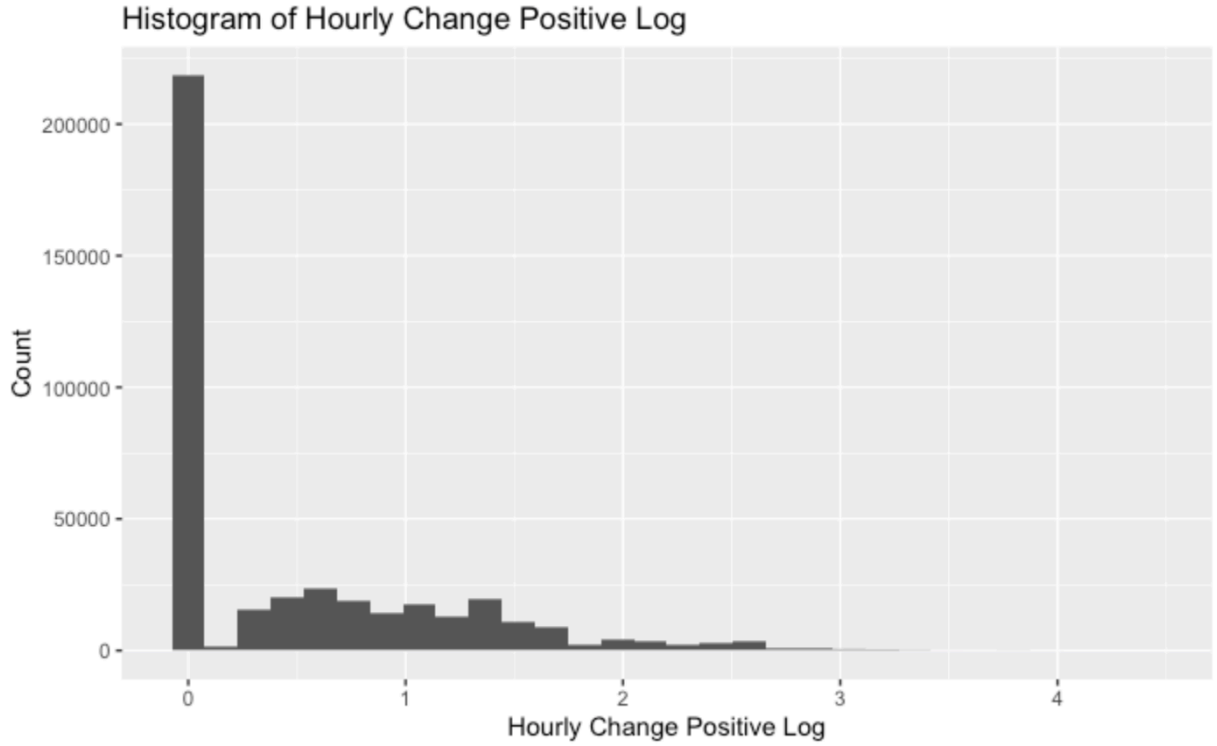
Appendix 3


Histogram of Hourly Change Positive
3e+05
2e+05
1e+05
0e+00
Count
0
20
40
60
80
Hourly Change Positive

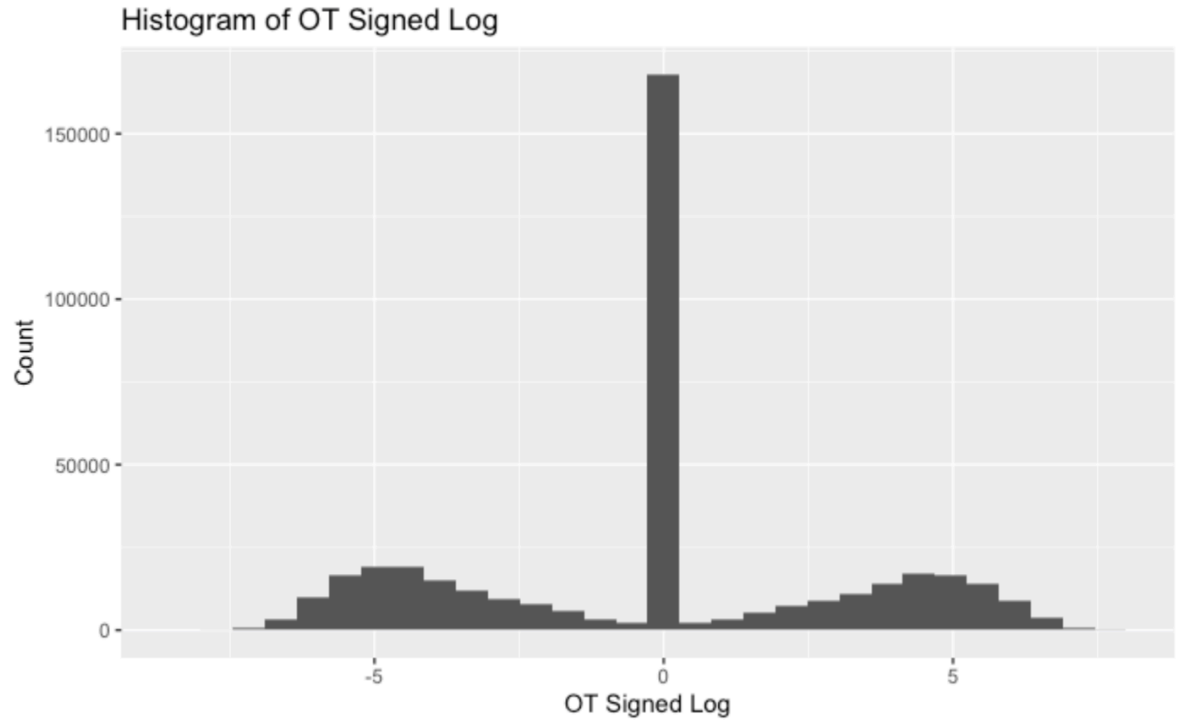
Appendix 4


Histogram of Year-Over-Year OT Hours Change
3e+05
2e+05
1e+05
0e+00
Count
-2000
0
2000
OT Hours Change

Appendix 5



Appendix 6



Appendix 7

	model	test_year	Accuracy	Precision	Recall
1	ranger_rf	2019	0.7818003	0.5755262	0.7461479
2	random_baseline	2019	0.5884524	0.2698060	0.3028321
3	ranger_rf	2020	0.8240475	0.5913027	0.7182154
4	random_baseline	2020	0.6117180	0.2267524	0.2976190
5	ranger_rf	2021	0.7721584	0.3694621	0.6818774
6	random_baseline	2021	0.6478135	0.1524029	0.2829205
7	ranger_rf	2022	0.7002013	0.2973091	0.5238607
8	random_baseline	2022	0.6474171	0.1748183	0.2731318
9	ranger_rf	2023	0.7594150	0.5743947	0.6721356
10	random_baseline	2023	0.6000522	0.2951546	0.2686928

Appendix 8

	Overall <dbl>
pay_basisper Hour	9899.318431
ot_pay_quintile	9456.504523
seniority	9299.163004
hourly_change_pos_log	4240.836324
ot_change_signed_log	2716.183205
agency_nameDEPT OF PARKS & RECREATION	2194.875533
agency_nameDEPARTMENT OF SANITATION	1162.187233
work_borough_locationQUEENS	965.043897
work_borough_locationBROOKLYN	224.958721
pay_basisper Day	218.650338