Dear Mayor Schewel,

I am excited to work with you on the new "affordable housing for all" tax that will be charged to all citizens making more than $50k/year. While it is concerning that you do not have income information about the residents of your city and want to use a predictive model to determine individuals' tax liability, I have developed a model using the 1996 Census data you provided to effectively predict whether individuals earn more or less than $50k per year.

In order to ensure predictive models produced using this dataset will generalize to broader populations, and provide results that will be easily interpretable by RTI and City of Durham staff, I made a few changes to the dataset before analysis:

1. The "relationships" data was dropped; It was redundant given marital status and gender were included.
2. Capital gains and capital losses were combined, with losses recorded as negative gains.
3. Country of origin was replaced with the income level of each country as gauged by World Bank per capita GDP figures.
4. Education levels were condensed into a smaller number of more interpretable levels (9 total).
5. For records missing 2 values or less, missing values were filled in using the field's median (for numerical fields) or mode (for text fields). Records missing > 2 values were dropped.

I tested two types of models on this data: a Logistic Regression classifier, along with a Random Forest classifier. Hundreds of hyper-parameter settings were tested for each model. The models "learned" from a subset of 60% of the census data. The models were then tested on a different subset of 20% of the Census data: the validation set. This method ensured we would not select a model so finely tuned to the training data set that it struggled to generalize to new, real-world data. The results achieved on the validation set by our top-performing Logistic Regression and Random Forest models are below:

Validation set results:

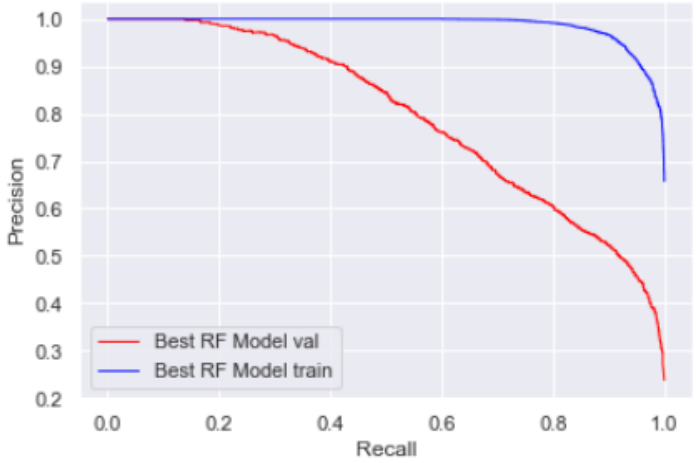
Model	Accuracy	Precision	Recall	F_β
Always under 50k	0.76	N/A	N/A	N/A
Logistic Regression	0.85	0.71	0.61	0.66
Random Forest	0.86	0.77	0.61	0.70

Accuracy is defined as the number of correct predictions divided by the total number of predictions made. While this may seem like the only metric needed, if our model simply predicted "under 50k" for every record in the dataset, it would achieve about 76% accuracy, as 76% of all subjects in the dataset earned under 50k. Metrics that place more emphasis on our model's ability to predict the rarer class (citizen's who make more than 50k) are needed. Precision, recall, and F-beta scores meet this need.

Those most concerned with ensuring we avoid saddling people who make under 50k with unnecessary taxes might suggest maximizing precision--the proportion of those we predict make over 50k that actually make over 50k. Those most concerned with minimizing the number of free-riders might suggest maximizing recall--the proportion of all those who actually make over 50k that we correctly identify as making over 50k. This would also ensure we maximize total tax revenues from those making over 50k.

The F-beta ($\boldsymbol{F_\beta}$) score allows us to balance precision and recall as we see fit. By setting β, we can define the relative importance of precision and recall. Setting $\beta < 1$ places more of an emphasis on precision than recall. Because you are likely more concerned by the possibility of accidentally hitting a low income

individual with an unfair tax than accidentally letting a higher income individual avoid the tax, I optimized the models tested to achieve the best possible F-beta score with $\beta = 0.8$. Keep in mind I could not set this score too much lower, or simply seek to optimize precision, as we are still seeking to maximize tax revenues as much as possible as well.



The plot above for our best performing random forest model illustrates the trade-off between precision and recall. The model is able to achieve >0.9 on both metrics on the test set, but when generalized to the validation set, balancing precision and recall becomes more challenging.

Among all models tested, the optimal Random Forest model achieved the highest F-beta score (0.70) while also outperforming the optimal Logistic Regression model on accuracy and precision. When run again on the final 20% of the data (the test set), the Random Forest classifier proved consistent, achieving an F-beta score of 0.69 and an accuracy of 0.86.

While a great deal more work needs to be done to ensure this tax is enacted effectively, and gathering actual income information would be preferable to predicting salaries, given the data available I feel confident that the Random Forest classifier is the best possible predictor of whether individual citizens of Durham make >50k.

Looking forward to our meeting Thursday,

David Henderson