Memo

To:	ManpowerGroup
From:	Rita Kurban
Date:	10/21/1979
Re:	Decision memo that evaluates the impact of the Lalonde program on people with and without a high school degree

Executive Summary

The purpose of this memorandum is to determine whether to target a job skills training program at individuals with high school degrees or individuals without high school degrees. To understand whether the Lalonde program had different impacts on these two subgroups, I interpreted the results from an observational study undertaken in 1978. After devising a linear regression model, running two random forests and implementing a Fisher Exact Test (FET), I concluded that if we had to choose a specific group, it would be better to target people with high school degrees because the predicted treatment effect is higher for this subgroup. However, the results produced by the models are not statistically significant for this particular dataset, and more data might be required. From the Fisher Exact Test, we know for sure that we can reject the sharp null hypothesis of no treatment effect for people with high school degrees which means the training program has a significant effect on at least some units.

Linear regression

To understand the impact of the high school degree on the income I ran one multiple linear regression on the subgroup with high school degrees and another one on the subgroup without degrees.

When devising a model, I first used all the independent variables available to predict the income. After calling the summary function, I identified the predictors with the highest p-values and eliminated them. A small p-value for the intercept and the slope allows us to conclude that there is a relationship between the independent and dependent variables. Consequently, if the p-value is high the relationship is not statistically significant which means that this independent variable doesn't improve the prediction. By removing the predictors with high p-values, I slightly improved both the adjusted R^2 (a measure of how well the model fits the actual data which adjusts for the number of variables considered) and the Residual Standard Error (RSE) which is a measure of the quality of a linear regression fit. However, even after these manipulations, the RSE remained high (7081 for "degree" and 6223 for "nodegree"). In the first model, for people with high school degree, the predicted coefficient for treatment is 2223. The pi-value for the treatment is around 14% which is not statistically significant. I estimated the confidence intervals and found out that in 95% of the cases the real value lies between -684.6609 and 5130.8773. This confidence interval suggests that the treatment can both increase and decrease the income which means that, for these data, we cannot claim a positive effect of the treatment on the income:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.822e+04  1.211e+04  -1.505   0.1359
treat        2.223e+03  1.484e+03   1.498   0.1375
educ         1.744e+03  1.002e+03   1.741   0.0851 .
black       -2.029e+03  1.859e+03  -1.092   0.2778
re74         2.427e-01  1.880e-01   1.291   0.2001
re75         6.936e-01  4.544e-01   1.526   0.1304
u75          3.753e+03  1.906e+03   1.969   0.0521 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7081 on 90 degrees of freedom
Multiple R-squared:  0.1771,    Adjusted R-squared:  0.1222
F-statistic: 3.227 on 6 and 90 DF,  p-value: 0.00644
```

For the second subgroup, I got a similar result. The estimated increase in the income is 1044. However, the p-value for this predictor is around 13%, which means that there is a 13% chance of getting such a result by chance. The confident interval is from -323.3571 to 2412.6447. It implies that it is impossible to claim a statistically significant positive treatment effect using this regression model:

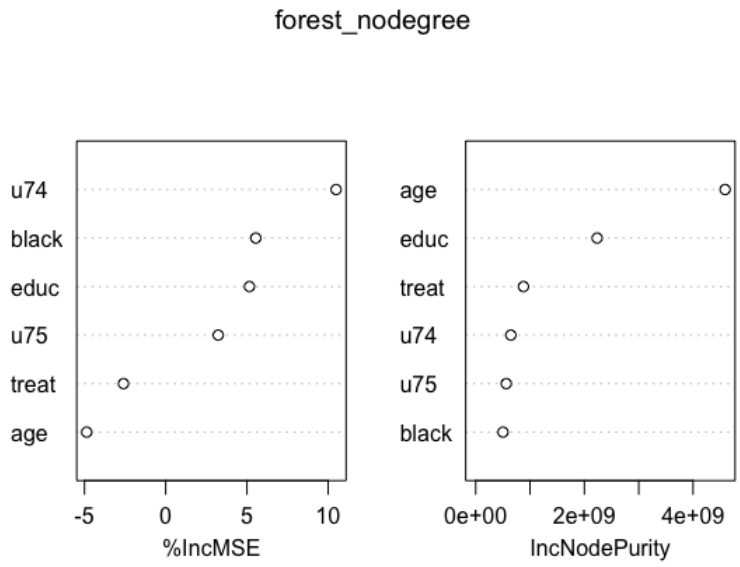
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3836.22    2701.03   1.420   0.1564
treat        1044.64     697.97   1.497   0.1354
age            48.49      46.44   1.044   0.2972
educ          215.77     225.94   0.955   0.3403
black       -2402.58     929.24  -2.586   0.0101 *
u74          1208.30    1089.45   1.109   0.2682
u75         -2174.55    1018.86  -2.134   0.0335 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6223 on 341 degrees of freedom
Multiple R-squared:  0.04451,   Adjusted R-squared:  0.0277
F-statistic: 2.648 on 6 and 341 DF,  p-value: 0.01596
```

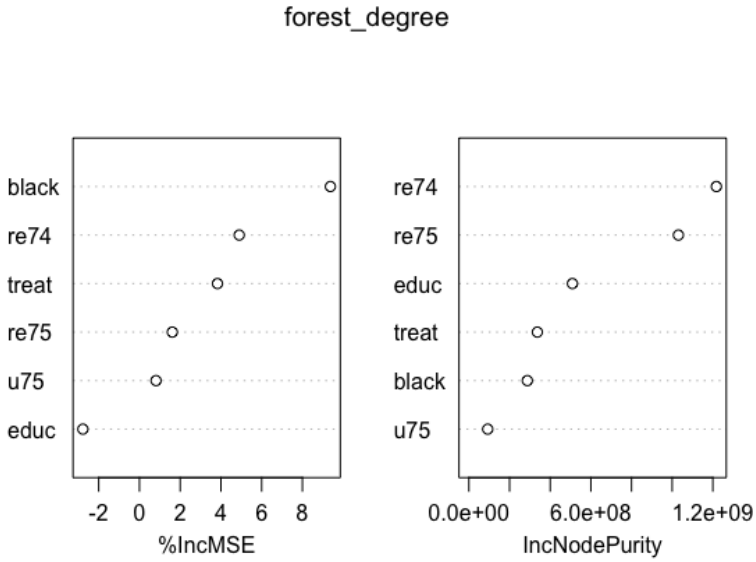
It is important to note that linear models assume a constant treatment effect, as there is one coefficient for the treatment variable, which means for any given input of predictor variables, we see the same treatment effect.

According to the models, the estimated treatment effect is higher for the subgroup with high school degrees. However, the data given is not enough to prove that this result is statistically significant. To further analyze the impact of the high school degree on the effectiveness of the treatment, I used a different model - Random Forest.

Random Forest

I created a random forest for high school grads and a separate random forest for non-high school grads using the same predictors as in (1) above. Then, using these models, I predicted counterfactual outcomes for the treated units by changing their treatment status to zero using the predict function. The resulting differential average treatment effect was 2260.7 and 742.6 for people with and without degrees respectively, which corresponds to the results the linear models produced. But how effective are these random forests? The Mean of Squared residuals is extremely high for both random forests (60947676 for degree and 47055903 for no degree). The variance explained is a negative number for both forests which suggests that even random variables would perform better than predictors that we used which significantly undermines the trustworthiness of the model. Moreover, the plots below demonstrate %IncMSE which is a very robust and informative measure. The higher it is, the better. The variable importance plots suggest that the treatment doesn't decrease the MSE much (3.8% for "degree"). In one of the plots (forest_nodegree) I even got a negative number which suggests that the fact that people without a high school degree received treatment does not have a statistically significant effect on the predicted outcome.



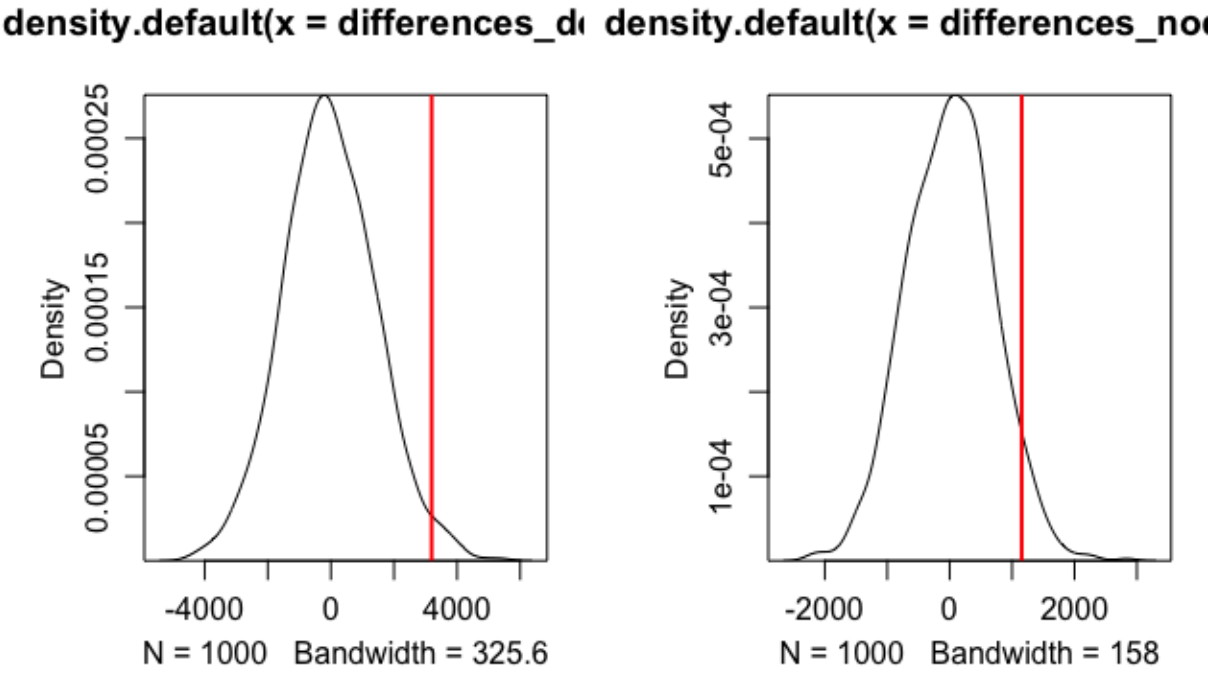


As opposed to the multiple regression models, the trees in the forest randomly choose different estimator variables. That is why forests can return different treatment effects for individual units. As a result, the treatment effect is not constant. To prove it, I calculated the treatment effect for individuals and found out that it differed significantly among units.

```
> treat_degree$re78 - forest_degree_counter
         3          7          9         10         12         17         19
22090.6468 -2818.8532 -3602.7982  1659.9476 12430.2517  9917.5517  7985.4468
        20         22         34         35         36         38         41
 7928.5468 -3181.9312 -2818.8532   372.8968 13774.1672  3092.6968 -1564.2732
        50         54         59         70         72         74         75
 2024.3268  -469.8832  6435.7672  2815.3117 -4017.1433  9146.9468  2914.1537
        77         81         82         83         84         89         94
15859.2468 -2818.8532  -497.7432 -5816.3024 -4664.3483 -2709.5524  7164.9268
        96         98        107        113        115        116        117
-5170.6524  -362.7032 29434.9517 -5098.5598 -4927.6211  4093.4030 14200.0178
       118        120        121        124        127        129        139
 -340.8624 -3611.5975 10877.3786 -4819.1136 -4376.0155 -5802.9316 -4136.4734
       142        145        152        154        156        158        165
-3792.0984 -5292.0620   824.2329  3254.7344 -4625.7809  3656.7413  2913.7103
       172        177        179        181        182
-2692.5410   422.6631  4026.8441  1557.2913 11231.5665
```

Fisher Exact Test

Last but not least, I implemented a Fisher Exact Test (FET) using the sharp hypothesis of no treatment effect for both subgroups. With the help of the experiment function, I randomly assigned individuals to either treatment or control. For each possible assignment, I calculated the value of the test statistic (the difference of means of income in 1978 for the treatment and control groups) that would have been observed under that assignment. After that, I compared these values to the observed treatment effect and identified how high the probability of getting equal or more extreme results is (p-value). For people with the degree, we can reject the sharp null hypothesis because the p-value is very small -- 2.1%(the plot on the left). It is highly unlikely to get such result by chance, that's why we can be almost sure that there is a non-zero treatment effect for at least some units. The p-value for people in the no-degree subgroup is around 5% (the plot on the right) which also implies that the probability of getting such a result purely by chance is low. However, it is higher than for people with a degree.



Recommendation

I used three methods to evaluate the impact of the Lalonde program on people with and without school degree. For the given dataset, the conclusions of the linear regressions and random forests were not statistically significant. However, both of them predicted that the positive treatment effect would be higher for the subset with high school degrees. The Fisher Exact Test suggests that there is a treatment effect for at least some units for both subgroups. More data are required to prove these findings.

A link to the code: https://gist.github.com/ritakurban/a51e705de3063fe32a6f9bb4ba842a7d