Income Project

Michelle Salandy

6/15/2019

1. Introduction

The Project performs exploratory data analysis and develops a machine learning algorithm to predict whether an individual will have an income earning greater than $50,000 annually utilizing data derived from a data-set on Kaggle.

Several predictive algorithms were trained including a GLM, KNN, Random Forest and rpart model using a series of continuous and categorical predictors or covariates to explain variances in earning potential. The accuracy of the models was utilized to evaluate performance and determine the 'best-fit' model.

The project has five sections. Following this brief introduction, the second section will outline the source of the data-set and the packages used. The third section will discuss the general properties of the data-set and the preprocessing of predictors. The fourth section will discuss the partitioning of the data-set and present a detailed discussion on the model fit and selection of the chosen algorithm. The report ends with a summary of the major findings and the identification of the accuracy of the '*best-fit*' algorithm.

2. Data-set and Packages

Load Data-set

This data-set used from Kaggle was originally extracted from the 1994 Census bureau database by Ronny Kohavi and Barry Becker (Data Mining and Visualization, Silicon Graphics). The following code is used to download the data-set.

```
url <- "https://github.com/msalandy/Income/blob/master/adultcsv.csv?raw=true"
salary<- read.csv(url)
```

Load Packages

The packages used to perform the analysis include:

- tidyverse
- caret
- corrplot
- gridExtra
- ggplot2
- dplyr
- data.table
- ROCR
- rpart
- randomForest

3. Data Exploration

General properties of the data-set

The income data-set has 32561 rows and 15 columns.

```
dim(salary)
```

The structure of this data-set indicates that missing values in the file has been represented by a question mark (?). These were replaced in the table with NA's and tallied, revealing the presence of several missing data points in Workclass (1836), occupation (1843), and native.country(583).

```
sapply(salary,function(x) sum(is.na(x)))
```

Dependent Variable

The categorical outcome to be predicted is income. It has two categories >50k and <=50k, with 24.1% of the entries representing individuals earning more than $50000 a year and 75.9% representing individuals earning less than $50000 a year.

Table 1. Category Totals

income	Count
<=50K	24720
>50K	7841

Independent Variables/A Priori Predictors

The a priori predictors provided in each column in the data-set are listed below.

1. Age: age
2. Employment status: workclass
3. Demographic weighting in the data-set: fnlwgt
4. Educational achievement: education
5. Numerical representing/ranking of educational achievement: education.num
6. Marital status: marital.status
7. Occupation: occupation
8. Familiar relationships: relationship
9. Race: race
10. Sex: sex
11. Capital gain: capital.gain
12. Capital loss: capital.loss
13. Hours of work per week: hours.per.week
14. Country of origin: native.country

Each predictor gives information regarding the characteristics of an individual. However, several predictors have similarities. For example, education.num provides a similar representation for educational achievement and thus one of the two can probably be omitted. Relationship is also omitted from the ongoing analysis as the two predictors marital.status and relationship share similar attributes, as a person's marital status describes a person's relationship with another.

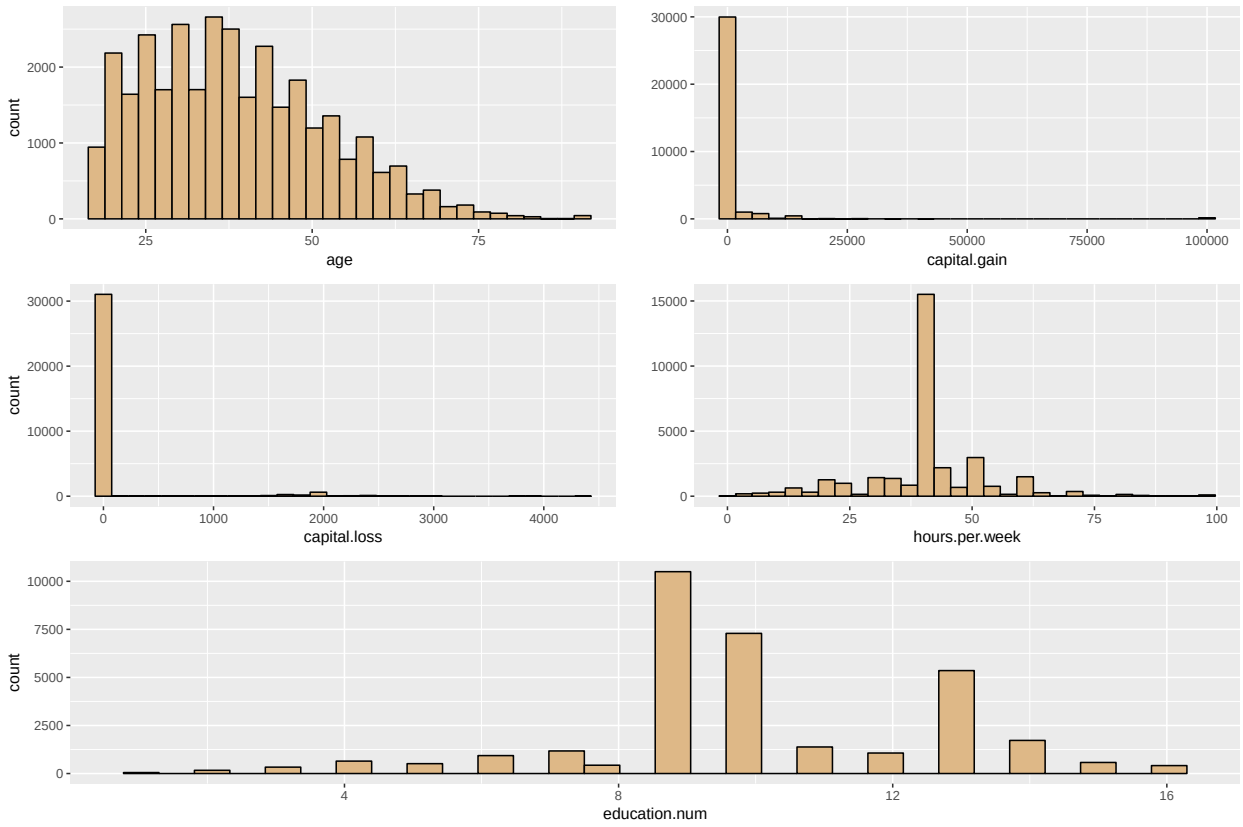
Several of the predictors also had various levels or unique categories. Table 1 shows that the demographic weigthing (fnlwgt)has 26,648 levels and as a result is also excluded from the analysis. It is removed because such a large number of levels would suggest that it will not group individuals into identifiable demographic categories that can significantly contribute to the prediction of the income of an individual.

Table 2. Unique Levels

	Unique Levels
age	73
Workclass	9
fnlwgt	21648
education	16
educationnum	16
maritalstatus	7
occupation	15
relationship	6
race	5
sex	2
capitalgain	119
capitalloss	92
hoursperweek	94
nativecountry	42

Further examination of the remaining continuous predictors by histogram plots displayed in Figure 1 shows that Capital gain and capital loss consists of mostly zeros.

Figure 1. Frequency of Continuous Predictors



Zeros account for 91.7% of the entries within the capital gain predictor and 95.3% zeroes in the capital loss predictor, which is a total of 29849 and 31042 entries from each of the respective predictors. The significant number so zeros in both a priori predictors are neither skewed to being related to a person earning <=50K or >50K and thus excluded from the ongoing analysis.

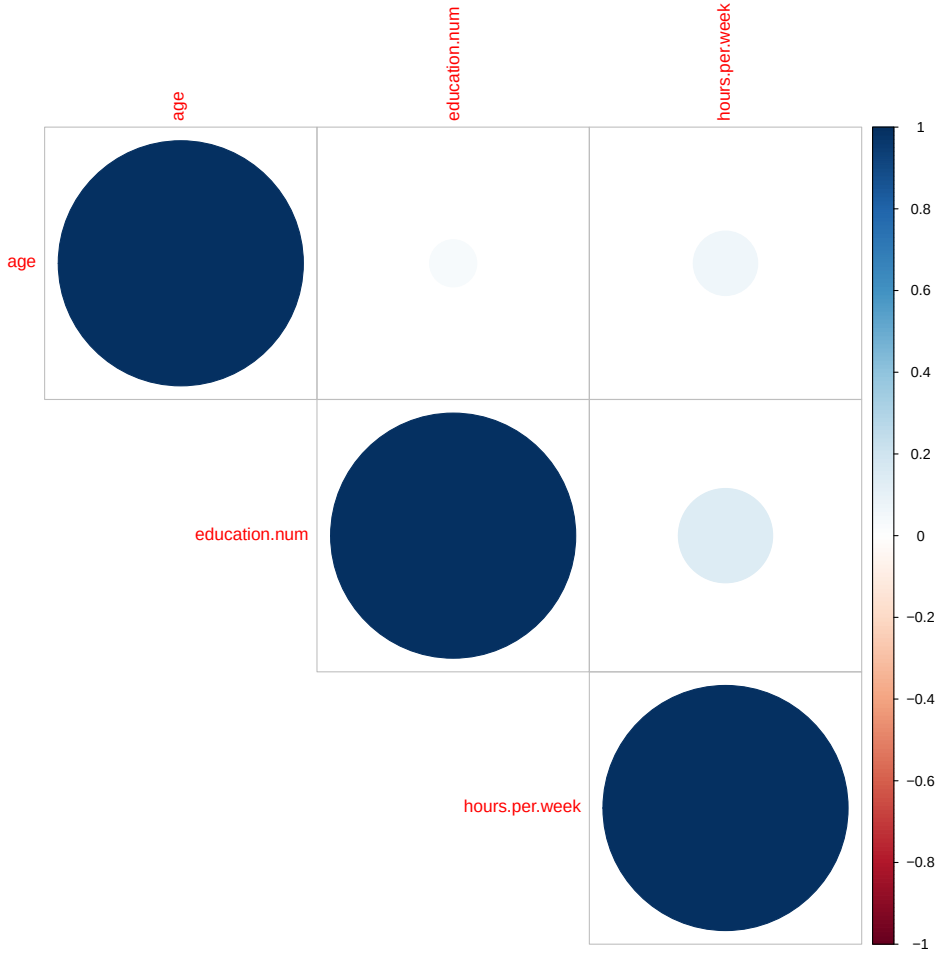
```
sum(salary$capital.gain==0)/length(salary$capital.gain)
sum(salary$capital.gain==0)
sum(salary$capital.loss==0)/length(salary$capital.loss)
sum(salary$capital.loss==0)
```

The correlation of the remaining a priori determinants age, education.num and hour.per.week which are displayed in Figure 2 show that they are not highly correlated.

Table 3. Correlation of Continuous Predictors

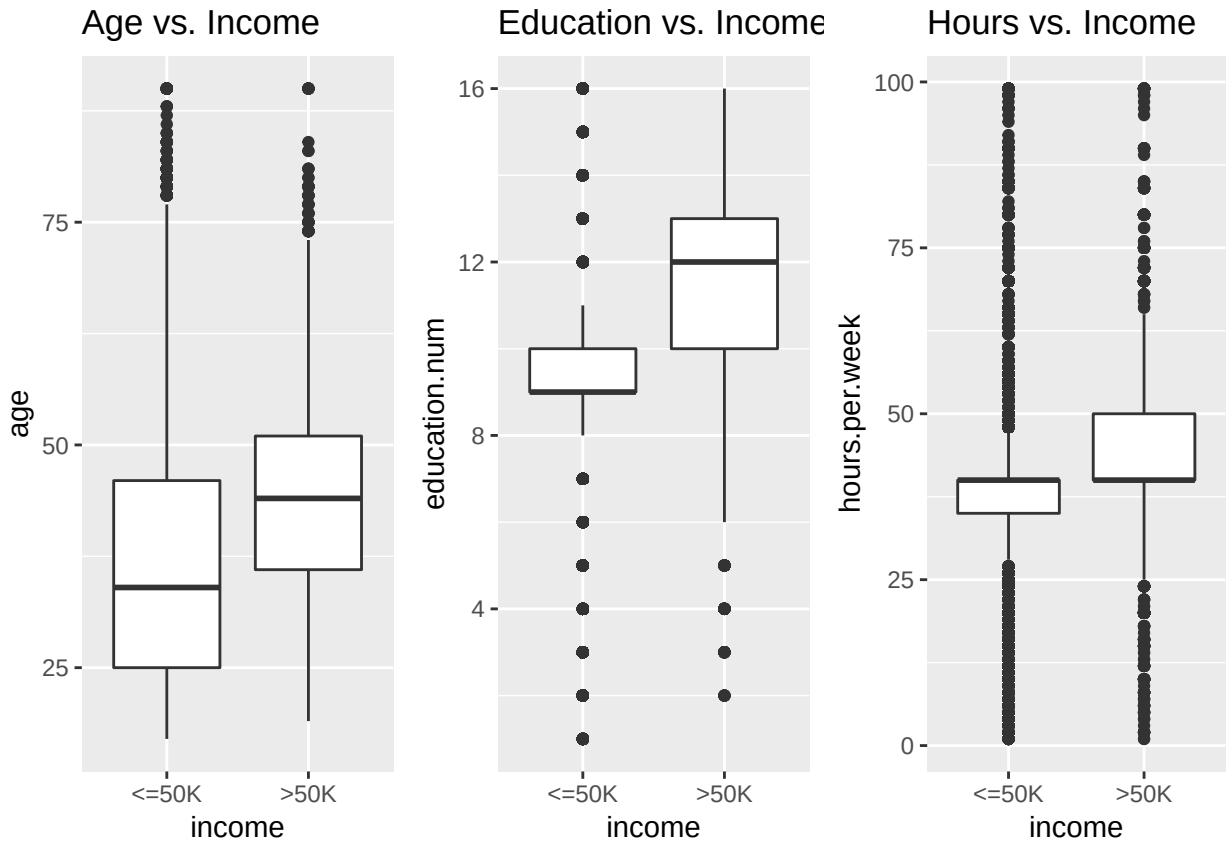
	age	education.num	hours.per.week
age	1.000	0.037	0.069
education.num	0.037	1.000	0.148
hours.per.week	0.069	0.148	1.000

Figure 2. Illustration of Correlation of Continuous Predictors



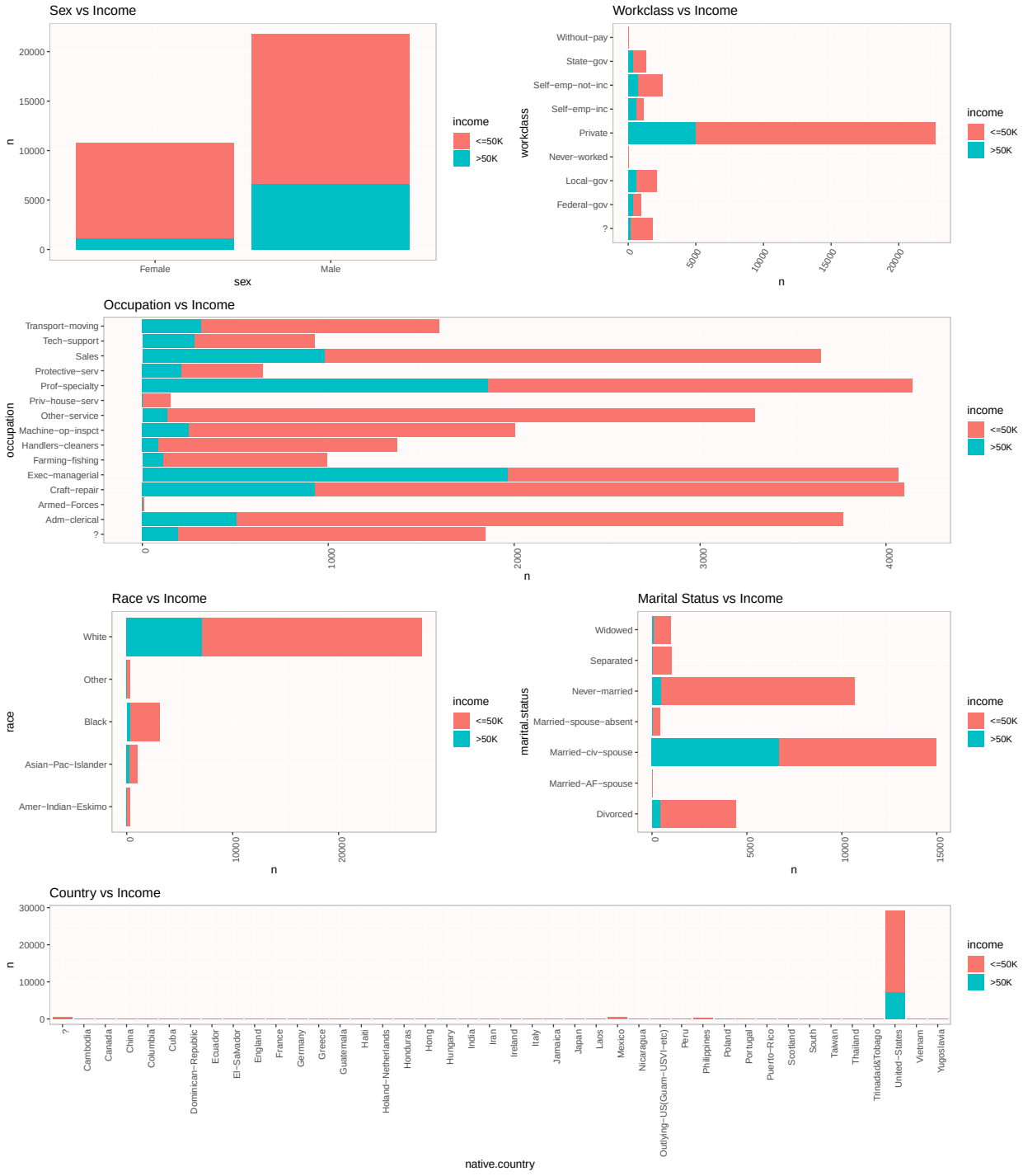
The three non-correlated continuous predictors show variations with income level. The box plots suggest that older individuals, those with higher educational achievements and working more than 40 hours per week have greater numbers earning more than $50000 annually (>50k) and can significantly influence variation.

Figure 3. Continuous Predictors and Income



Exploration of the categorical predictors indicates that one predictor (native.country) does not have significant variation and can be excluded from the ongoing analysis. The histograms also suggest that white males in the private job class with a professional specialty or in executive managerial positions whom are married-civ spouse have greater numbers earning more than $50000 annually (>50k) and can significantly influence variation.

Figure 4. Categorical Predictors and Income



4. Preprocessing and Model Fit Results

Removing predictor columns

Considering the findings after the analysis of the raw data-set,it was streamlined and several predictors were removed. These included fnlwgt, education, relationship, capital.gain, capital.loss, native.country and rows including missing data.

Data Partitioning

The categorical dependent variable is transformed to a binary factor with 0 representing an earning of <=50K and 1 representing an earning of >50K and then the data-set was then split into a training set denoted as train and a validation set denoted as test, which is used to test the algorithm for predicting income.The test set represents 10% of the data-set.

```
set.seed(1)
test_index <- createDataPartition(y = salary$income, times = 1, p = 0.10, list = FALSE)
train <- salary %>% slice(-test_index)
test <- salary %>% slice(test_index)
```

Fit Model

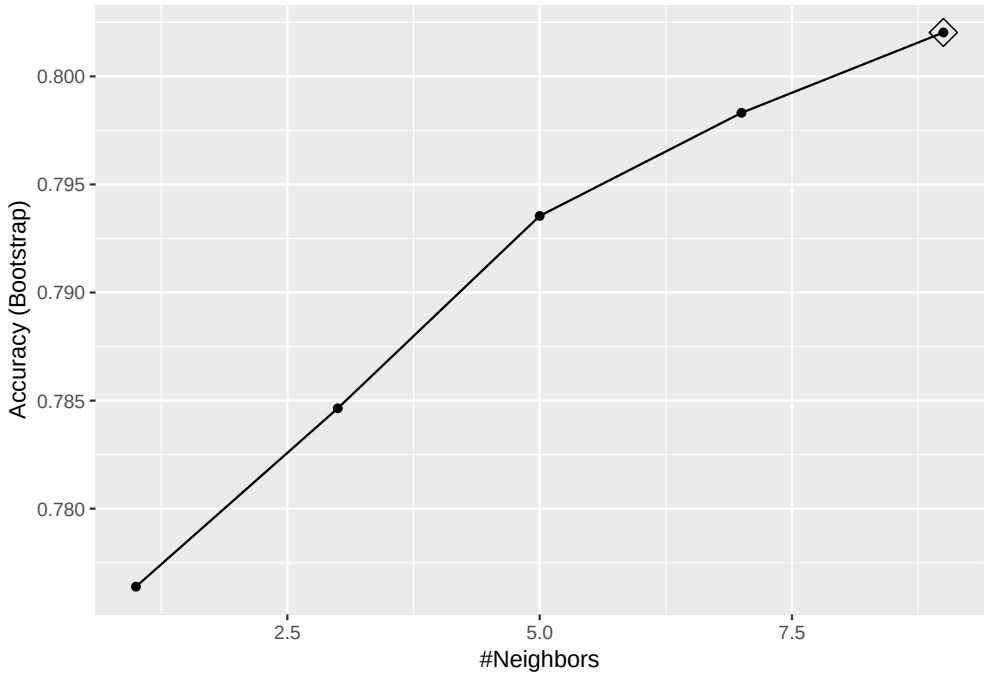
- GLM: Logistic Regression

```
fitglm <-glm(income ~.,family=binomial(link='logit'),data=train)
phat<- predict(fitglm, test, type = "response")
yhat<- ifelse(phat>0.5, 1,0) %>% factor()
confusionMatrix(yhat, test$income)$overall["Accuracy"]
```

```
##  Accuracy
## 0.8378876
```

- KNN: k-Nearest Neighbor

```
fitknn <- train(income ~ ., method = "knn", tuneGrid = data.frame(k = seq(1, 10, 2)), data = train)
ggplot(fitknn, highlight = TRUE)
```




```
fitknn$bestTune
```

```
##   k
## 5 9
```

```
fitknn <-knn3(income ~.,data=train, k=9)
yhatknn<-predict(fitknn, test, type="class")
confusionMatrix(yhatknn, test$income)$overall["Accuracy"]
```

```
##  Accuracy
## 0.8182376
```

- Recursive Partitioning and Regression Trees: RPART

```
fitrpart<-train(income~., method="rpart", data=train)
yhat<-predict(fitrpart, test, type="raw")
confusionMatrix(yhat, test$income)$overall["Accuracy"]
```

```
##  Accuracy
## 0.8210009
```

- Random Forest

```
fitrf<- randomForest(income~.,data= train)
yhat<-predict(fitrf, test, type="class")
confusionMatrix(yhat, test$income)$overall["Accuracy"]
```

```
## Accuracy
## 0.841879
```

5.Conclusion

Using data from a Adult Income Census data-set sourced from Kaggle several predictors or covariates were utilized to predict income earnings of an individual. After the use of several models the highest accuracy of 0.841879 was established by a Random Forest model with predictors age, workclass, education, occupation, race, sex, hours of work per week and marital status.

Link to Github Movielens Project