finding_donors_updated_v2

January 29, 2026

0.1 Supervised Learning

0.2 Project: Finding Donors for *CharityML*

In this notebook, some template code has already been provided for you, and it will be your job to implement the additional functionality necessary to successfully complete this project. Sections that begin with **'Implementation'** in the header indicate that the following block of code will require additional functionality which you must provide. Instructions will be provided for each section and the specifics of the implementation are marked in the code block with a `'TODO'` statement. Please be sure to read the instructions carefully!

In addition to implementing code, there will be questions that you must answer which relate to the project and your implementation. Each section where you will answer a question is preceded by a **'Question X'** header. Carefully read each question and provide thorough answers in the following text boxes that begin with **'Answer:'**. Your project submission will be evaluated based on your answers to each of the questions and the implementation you provide.

> **Note:** Please specify WHICH VERSION OF PYTHON you are using when submitting this notebook. Code and Markdown cells can be executed using the **Shift + Enter** keyboard shortcut. In addition, Markdown cells can be edited by typically double-clicking the cell to enter edit mode.

0.3 Getting Started

In this project, you will employ several supervised algorithms of your choice to accurately model individuals' income using data collected from the 1994 U.S. Census. You will then choose the best candidate algorithm from preliminary results and further optimize this algorithm to best model the data. Your goal with this implementation is to construct a model that accurately predicts whether an individual makes more than $50,000. This sort of task can arise in a non-profit setting, where organizations survive on donations. Understanding an individual's income can help a non-profit better understand how large of a donation to request, or whether or not they should reach out to begin with. While it can be difficult to determine an individual's general income bracket directly from public sources, we can (as we will see) infer this value from other publically available features.

The dataset for this project originates from the UCI Machine Learning Repository. The datset was donated by Ron Kohavi and Barry Becker, after being published in the article *"Scaling Up the Accuracy of Naive-Bayes Classifiers: A Decision-Tree Hybrid"*. You can find the article by Ron Kohavi online. The data we investigate here consists of small changes to the original dataset, such as removing the `'fnlwgt'` feature and records with missing or ill-formatted entries.

0.4 Exploring the Data

Run the code cell below to load necessary Python libraries and load the census data. Note that the last column from this dataset, `'income'`, will be our target label (whether an individual makes more than, or at most, $50,000 annually). All other columns are features about each individual in the census database.

[1]:
```
# Import libraries necessary for this project
import numpy as np
import pandas as pd
from time import time
from IPython.display import display # Allows the use of display() for DataFrames

# Import supplementary visualization code visuals.py
import visuals_updated_v2 as vs

# Pretty display for notebooks
%matplotlib inline

# Load the Census dataset
data = pd.read_csv("census.csv")

# Success - Display the first record
display(data.head(10))
```

```
   age          workclass education_level  education-num  \
0   39          State-gov       Bachelors           13.0
1   50   Self-emp-not-inc       Bachelors           13.0
2   38            Private         HS-grad            9.0
3   53            Private            11th            7.0
4   28            Private       Bachelors           13.0
5   37            Private         Masters           14.0
6   49            Private             9th            5.0
7   52   Self-emp-not-inc         HS-grad            9.0
8   31            Private         Masters           14.0
9   42            Private       Bachelors           13.0

           marital-status          occupation    relationship    race  \
0           Never-married        Adm-clerical   Not-in-family   White
1      Married-civ-spouse     Exec-managerial         Husband   White
2                Divorced   Handlers-cleaners   Not-in-family   White
3      Married-civ-spouse   Handlers-cleaners         Husband   Black
4      Married-civ-spouse      Prof-specialty            Wife   Black
5      Married-civ-spouse     Exec-managerial            Wife   White
6   Married-spouse-absent       Other-service   Not-in-family   Black
7      Married-civ-spouse     Exec-managerial         Husband   White
8           Never-married      Prof-specialty   Not-in-family   White
9      Married-civ-spouse     Exec-managerial         Husband   White
```

	sex	capital-gain	capital-loss	hours-per-week	native-country	income
0	Male	2174.0	0.0	40.0	United-States	<=50K
1	Male	0.0	0.0	13.0	United-States	<=50K
2	Male	0.0	0.0	40.0	United-States	<=50K
3	Male	0.0	0.0	40.0	United-States	<=50K
4	Female	0.0	0.0	40.0	Cuba	<=50K
5	Female	0.0	0.0	40.0	United-States	<=50K
6	Female	0.0	0.0	16.0	Jamaica	<=50K
7	Male	0.0	0.0	45.0	United-States	>50K
8	Female	14084.0	0.0	50.0	United-States	>50K
9	Male	5178.0	0.0	40.0	United-States	>50K

0.4.1 Implementation: Data Exploration

A cursory investigation of the dataset will determine how many individuals fit into either group, and will tell us about the percentage of these individuals making more than $50,000. In the code cell below, you will need to compute the following: - The total number of records, `'n_records'` - The number of individuals making more than $50,000 annually, `'n_greater_50k'`. - The number of individuals making at most $50,000 annually, `'n_at_most_50k'`. - The percentage of individuals making more than $50,000 annually, `'greater_percent'`.

** HINT: ** You may need to look at the table above to understand how the `'income'` entries are formatted.

[2]:
```
# TODO: Total number of records
n_records = data.shape[0]

# TODO: Number of records where individual's income is more than $50,000
n_greater_50k = data[data['income'] == '>50K'].shape[0]

# TODO: Number of records where individual's income is at most $50,000
n_at_most_50k = data[data['income'] == '<=50K'].shape[0]

# TODO: Percentage of individuals whose income is more than $50,000
greater_percent = n_greater_50k/n_records*100

# Print the results
print("Total number of records: {}".format(n_records))
print("Individuals making more than $50,000: {}".format(n_greater_50k))
print("Individuals making at most $50,000: {}".format(n_at_most_50k))
print("Percentage of individuals making more than $50,000: {:.2f}%".
  ↪format(greater_percent))
```

```
Total number of records: 45222
Individuals making more than $50,000: 11208
Individuals making at most $50,000: 34014
Percentage of individuals making more than $50,000: 24.78%
```

** Featureset Exploration **

- **age**: continuous.
- **workclass**: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
- **education**: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
- **education-num**: continuous.
- **marital-status**: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
- **occupation**: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.
- **relationship**: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
- **race**: Black, White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other.
- **sex**: Female, Male.
- **capital-gain**: continuous.
- **capital-loss**: continuous.
- **hours-per-week**: continuous.
- **native-country**: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.

0.5 Preparing the Data

Before data can be used as input for machine learning algorithms, it often must be cleaned, formatted, and restructured — this is typically known as **preprocessing**. Fortunately, for this dataset, there are no invalid or missing entries we must deal with, however, there are some qualities about certain features that must be adjusted. This preprocessing can help tremendously with the outcome and predictive power of nearly all learning algorithms.

0.5.1 Transforming Skewed Continuous Features

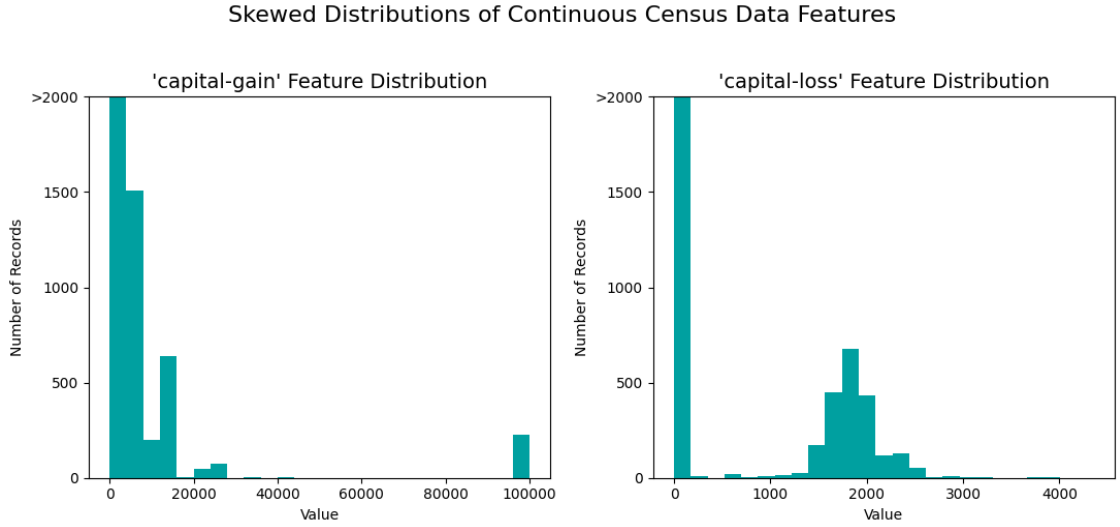
A dataset may sometimes contain at least one feature whose values tend to lie near a single number, but will also have a non-trivial number of vastly larger or smaller values than that single number. Algorithms can be sensitive to such distributions of values and can underperform if the range is not properly normalized. With the census dataset two features fit this description: '`capital-gain`' and `'capital-loss'`.

Run the code cell below to plot a histogram of these two features. Note the range of the values present and how they are distributed.

[3]:
```
# Split the data into features and target label
income_raw = data['income']
features_raw = data.drop('income', axis = 1)
```

```
# Visualize skewed continuous features of original data
vs.distribution(data)
```

```
c:\Users\Cristi\Python_Projects\UdaCity ML\Project1
Supervised_Learning\visuals_updated_v2.py:48: UserWarning: FigureCanvasAgg is
non-interactive, and thus cannot be shown
  fig.show()
```



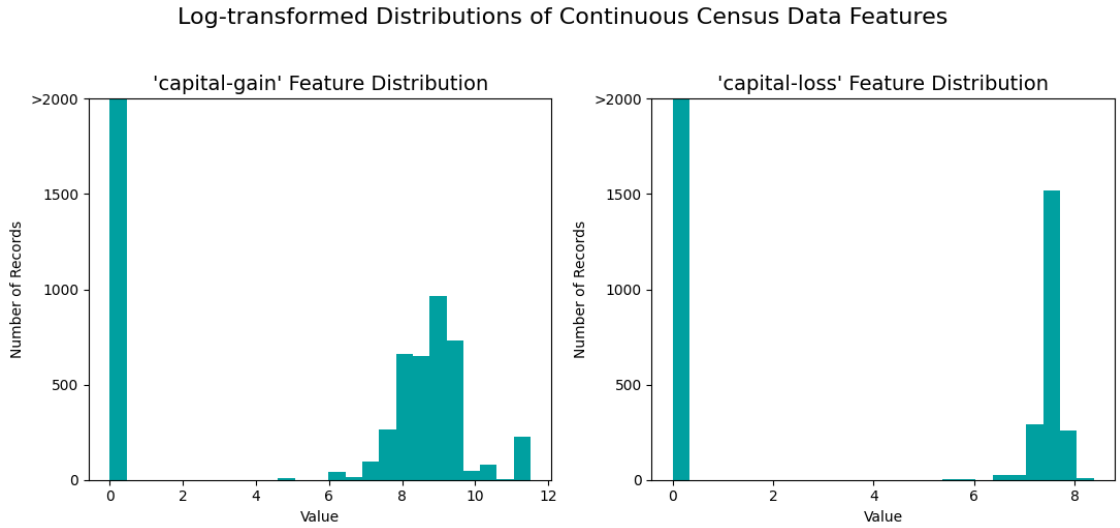
For highly-skewed feature distributions such as `'capital-gain'` and `'capital-loss'`, it is common practice to apply a logarithmic transformation on the data so that the very large and very small values do not negatively affect the performance of a learning algorithm. Using a logarithmic transformation significantly reduces the range of values caused by outliers. Care must be taken when applying this transformation however: The logarithm of `0` is undefined, so we must translate the values by a small amount above `0` to apply the the logarithm successfully.

Run the code cell below to perform a transformation on the data and visualize the results. Again, note the range of values and how they are distributed.

[4]:
```
# Log-transform the skewed features
skewed = ['capital-gain', 'capital-loss']
features_log_transformed = pd.DataFrame(data = features_raw)
features_log_transformed[skewed] = features_raw[skewed].apply(lambda x: np.
 ↪log(x + 1))

# Visualize the new log distributions
vs.distribution(features_log_transformed, transformed = True)
```

```
c:\Users\Cristi\Python_Projects\UdaCity ML\Project1
Supervised_Learning\visuals_updated_v2.py:48: UserWarning: FigureCanvasAgg is
non-interactive, and thus cannot be shown
  fig.show()
```



0.5.2 Normalizing Numerical Features

In addition to performing transformations on features that are highly skewed, it is often good practice to perform some type of scaling on numerical features. Applying a scaling to the data does not change the shape of each feature's distribution (such as `'capital-gain'` or `'capital-loss'` above); however, normalization ensures that each feature is treated equally when applying supervised learners. Note that once scaling is applied, observing the data in its raw form will no longer have the same original meaning, as exampled below.

Run the code cell below to normalize each numerical feature. We will use `sklearn.preprocessing.MinMaxScaler` for this.

[5]:
```
# Import sklearn.preprocessing.StandardScaler
from sklearn.preprocessing import MinMaxScaler

# Initialize a scaler, then apply it to the features
scaler = MinMaxScaler() # default=(0, 1)
numerical = ['age', 'education-num', 'capital-gain', 'capital-loss',␣
 ↪'hours-per-week']

features_log_minmax_transform = pd.DataFrame(data = features_log_transformed)
features_log_minmax_transform[numerical] = scaler.
 ↪fit_transform(features_log_transformed[numerical])

# Show an example of a record with scaling applied
display(features_log_minmax_transform.head())
```

```
        age          workclass education_level  education-num  \
0  0.301370          State-gov       Bachelors       0.800000
1  0.452055   Self-emp-not-inc       Bachelors       0.800000
```

```
2  0.287671           Private         HS-grad       0.533333
3  0.493151           Private            11th       0.400000
4  0.150685           Private       Bachelors       0.800000

       marital-status          occupation    relationship    race      sex  \
0       Never-married        Adm-clerical   Not-in-family   White     Male
1  Married-civ-spouse     Exec-managerial         Husband   White     Male
2            Divorced   Handlers-cleaners   Not-in-family   White     Male
3  Married-civ-spouse   Handlers-cleaners         Husband   Black     Male
4  Married-civ-spouse      Prof-specialty            Wife   Black   Female

   capital-gain  capital-loss  hours-per-week  native-country
0      0.667492           0.0        0.397959   United-States
1      0.000000           0.0        0.122449   United-States
2      0.000000           0.0        0.397959   United-States
3      0.000000           0.0        0.397959   United-States
4      0.000000           0.0        0.397959            Cuba
```

0.5.3 Implementation: Data Preprocessing

From the table in **Exploring the Data** above, we can see there are several features for each record that are non-numeric. Typically, learning algorithms expect input to be numeric, which requires that non-numeric features (called *categorical variables*) be converted. One popular way to convert categorical variables is by using the **one-hot encoding** scheme. One-hot encoding creates a *"dummy"* variable for each possible category of each non-numeric feature. For example, assume `someFeature` has three possible entries: `A`, `B`, or `C`. We then encode this feature into `someFeature_A`, `someFeature_B` and `someFeature_C`.

| someFeature | | someFeature_A | someFeature_B | someFeature_C |
:-: | :-: | | :-: | :-: | :-: |
0 | B | | 0 | 1 | 0 |
1 | C | —-> one-hot encode —-> | 0 | 0 | 1 |
2 | A | | 1 | 0 | 0 |

Additionally, as with the non-numeric features, we need to convert the non-numeric target label, `'income'` to numerical values for the learning algorithm to work. Since there are only two possible categories for this label ("<=50K" and ">50K"), we can avoid using one-hot encoding and simply encode these two categories as `0` and `1`, respectively. In code cell below, you will need to implement the following: - Use `pandas.get_dummies()` to perform one-hot encoding on the `'features_log_minmax_transform'` data. - Convert the target label `'income_raw'` to numerical entries. - Set records with "<=50K" to `0` and records with ">50K" to `1`.

[6]:
```
from IPython.display import display, HTML

# TODO: One-hot encode the 'features_log_minmax_transform' data using pandas.
  ↪get_dummies()
features_final = pd.get_dummies(features_log_minmax_transform)
#display(features_final.head())
display(HTML(features_final.head().to_html()))
```

```
# TODO: Encode the 'income_raw' data to numerical values
income = income_raw.map({'<=50K': 0, '>50K': 1})
display(income)

# Print the number of features after one-hot encoding
encoded = list(features_final.columns)
print("{} total features after one-hot encoding.".format(len(encoded)))

# Uncomment the following line to see the encoded feature names
print (encoded)
```

```
<IPython.core.display.HTML object>
```

```
0        0
1        0
2        0
3        0
4        0
        ..
45217    0
45218    0
45219    0
45220    0
45221    1
Name: income, Length: 45222, dtype: int64
```

```
103 total features after one-hot encoding.
['age', 'education-num', 'capital-gain', 'capital-loss', 'hours-per-week',
'workclass_ Federal-gov', 'workclass_ Local-gov', 'workclass_ Private',
'workclass_ Self-emp-inc', 'workclass_ Self-emp-not-inc', 'workclass_ State-
gov', 'workclass_ Without-pay', 'education_level_ 10th', 'education_level_
11th', 'education_level_ 12th', 'education_level_ 1st-4th', 'education_level_
5th-6th', 'education_level_ 7th-8th', 'education_level_ 9th', 'education_level_
Assoc-acdm', 'education_level_ Assoc-voc', 'education_level_ Bachelors',
'education_level_ Doctorate', 'education_level_ HS-grad', 'education_level_
Masters', 'education_level_ Preschool', 'education_level_ Prof-school',
'education_level_ Some-college', 'marital-status_ Divorced', 'marital-status_
Married-AF-spouse', 'marital-status_ Married-civ-spouse', 'marital-status_
Married-spouse-absent', 'marital-status_ Never-married', 'marital-status_
Separated', 'marital-status_ Widowed', 'occupation_ Adm-clerical', 'occupation_
Armed-Forces', 'occupation_ Craft-repair', 'occupation_ Exec-managerial',
'occupation_ Farming-fishing', 'occupation_ Handlers-cleaners', 'occupation_
Machine-op-inspct', 'occupation_ Other-service', 'occupation_ Priv-house-serv',
'occupation_ Prof-specialty', 'occupation_ Protective-serv', 'occupation_
Sales', 'occupation_ Tech-support', 'occupation_ Transport-moving',
'relationship_ Husband', 'relationship_ Not-in-family', 'relationship_ Other-
relative', 'relationship_ Own-child', 'relationship_ Unmarried', 'relationship_
```

```
Wife', 'race_ Amer-Indian-Eskimo', 'race_ Asian-Pac-Islander', 'race_ Black',
'race_ Other', 'race_ White', 'sex_ Female', 'sex_ Male', 'native-country_
Cambodia', 'native-country_ Canada', 'native-country_ China', 'native-country_
Columbia', 'native-country_ Cuba', 'native-country_ Dominican-Republic',
'native-country_ Ecuador', 'native-country_ El-Salvador', 'native-country_
England', 'native-country_ France', 'native-country_ Germany', 'native-country_
Greece', 'native-country_ Guatemala', 'native-country_ Haiti', 'native-country_
Holand-Netherlands', 'native-country_ Honduras', 'native-country_ Hong',
'native-country_ Hungary', 'native-country_ India', 'native-country_ Iran',
'native-country_ Ireland', 'native-country_ Italy', 'native-country_ Jamaica',
'native-country_ Japan', 'native-country_ Laos', 'native-country_ Mexico',
'native-country_ Nicaragua', 'native-country_ Outlying-US(Guam-USVI-etc)',
'native-country_ Peru', 'native-country_ Philippines', 'native-country_ Poland',
'native-country_ Portugal', 'native-country_ Puerto-Rico', 'native-country_
Scotland', 'native-country_ South', 'native-country_ Taiwan', 'native-country_
Thailand', 'native-country_ Trinadad&Tobago', 'native-country_ United-States',
'native-country_ Vietnam', 'native-country_ Yugoslavia']
```

0.5.4 Shuffle and Split Data

Now all *categorical variables* have been converted into numerical features, and all numerical features have been normalized. As always, we will now split the data (both features and their labels) into training and test sets. 80% of the data will be used for training and 20% for testing.

Run the code cell below to perform this split.

[7]:
```
# Import train_test_split
from sklearn.model_selection import train_test_split

# Split the 'features' and 'income' data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(features_final,
                                                    income,
                                                    test_size = 0.2,
                                                    random_state = 0)

# Show the results of the split
print("Training set has {} samples.".format(X_train.shape[0]))
print("Testing set has {} samples.".format(X_test.shape[0]))
```

```
Training set has 36177 samples.
Testing set has 9045 samples.
```

Note: this Workspace is running on ***sklearn*** *v0.19. If you use the newer version (>="0.20"), the* ***sklearn.cross_validation*** *has been replaced with* ***sklearn.model_selection****.*

0.6 Evaluating Model Performance

In this section, we will investigate four different algorithms, and determine which is best at modeling the data. Three of these algorithms will be supervised learners of your choice, and the fourth algorithm is known as a *naive predictor*.

0.6.1 Metrics and the Naive Predictor

CharityML, equipped with their research, knows individuals that make more than $50,000 are most likely to donate to their charity. Because of this, *CharityML* is particularly interested in predicting who makes more than $50,000 accurately. It would seem that using **accuracy** as a metric for evaluating a particular model's performace would be appropriate. Additionally, identifying someone that *does not* make more than $50,000 as someone who does would be detrimental to *CharityML*, since they are looking to find individuals willing to donate. Therefore, a model's ability to precisely predict those that make more than $50,000 is *more important* than the model's ability to **recall** those individuals. We can use **F-beta score** as a metric that considers both precision and recall:

$$F_\beta = (1 + \beta^2) \cdot \frac{precision \cdot recall}{(\beta^2 \cdot precision) + recall}$$

In particular, when $\beta = 0.5$, more emphasis is placed on precision. This is called the $\mathbf{F}_{0.5}$ **score** (or F-score for simplicity).

Looking at the distribution of classes (those who make at most $50,000, and those who make more), it's clear most individuals do not make more than $50,000. This can greatly affect **accuracy**, since we could simply say *"this person does not make more than $50,000"* and generally be right, without ever looking at the data! Making such a statement would be called **naive**, since we have not considered any information to substantiate the claim. It is always important to consider the *naive prediction* for your data, to help establish a benchmark for whether a model is performing well. That been said, using that prediction would be pointless: If we predicted all people made less than $50,000, *CharityML* would identify no one as donors.

Note: Recap of accuracy, precision, recall ** Accuracy ** measures how often the classifier makes the correct prediction. It's the ratio of the number of correct predictions to the total number of predictions (the number of test data points).

** Precision ** tells us what proportion of messages we classified as spam, actually were spam. It is a ratio of true positives(words classified as spam, and which are actually spam) to all positives(all words classified as spam, irrespective of whether that was the correct classificatio), in other words it is the ratio of

`[True Positives/(True Positives + False Positives)]`

** Recall(sensitivity)** tells us what proportion of messages that actually were spam were classified by us as spam. It is a ratio of true positives(words classified as spam, and which are actually spam) to all the words that were actually spam, in other words it is the ratio of

`[True Positives/(True Positives + False Negatives)]`

For classification problems that are skewed in their classification distributions like in our case, for example if we had a 100 text messages and only 2 were spam and the rest 98 weren't, accuracy

by itself is not a very good metric. We could classify 90 messages as not spam(including the 2 that were spam but we classify them as not spam, hence they would be false negatives) and 10 as spam(all 10 false positives) and still get a reasonably good accuracy score. For such cases, precision and recall come in very handy. These two metrics can be combined to get the F1 score, which is weighted average(harmonic mean) of the precision and recall scores. This score can range from 0 to 1, with 1 being the best possible F1 score(we take the harmonic mean as we are dealing with ratios).

0.6.2 Question 1 - Naive Predictor Performace

- If we chose a model that always predicted an individual made more than $50,000, what would that model's accuracy and F-score be on this dataset? You must use the code cell below and assign your results to `'accuracy'` and `'fscore'` to be used later.

** Please note ** that the the purpose of generating a naive predictor is simply to show what a base model without any intelligence would look like. In the real world, ideally your base model would be either the results of a previous model or could be based on a research paper upon which you are looking to improve. When there is no benchmark model set, getting a result better than random choice is a place you could start from.

** HINT: **

- When we have a model that always predicts '1' (i.e. the individual makes more than 50k) then our model will have no True Negatives(TN) or False Negatives(FN) as we are not making any negative('0' value) predictions. Therefore our Accuracy in this case becomes the same as our Precision(True Positives/(True Positives + False Positives)) as every prediction that we have made with value '1' that should have '0' becomes a False Positive; therefore our denominator in this case is the total number of records we have in total.
- Our Recall score(True Positives/(True Positives + False Negatives)) in this setting becomes 1 as we have no False Negatives.

[8]:
```
TP = np.sum(income) # Counting the ones as this is the naive case.
# Note that 'income' is the 'income_raw' data encoded to numerical values done␣
 ↪in the data preprocessing step.
FP = income.count() - TP # Specific to the naive case

TN = 0 # No predicted negatives in the naive case
FN = 0 # No predicted negatives in the naive case
print(TP, FP, TN, FN)

# TODO: Calculate accuracy, precision and recall

accuracy = (TP+TN)/(TP + TN + FP + FN)
recall = TP/(TP + FN)
precision = TP/(TP + FP)

# TODO: Calculate F-score using the formula above for beta = 0.5 and correct␣
 ↪values for precision and recall.
beta = 0.5
```

```
fscore = (1 + beta**2) * precision * recall / (beta**2 * precision + recall)

# Print the results
print("Naive Predictor: [Accuracy score: {:.4f}, F-score: {:.4f}]".
  ↪format(accuracy, fscore))
```

```
11208 34014 0 0
Naive Predictor: [Accuracy score: 0.2478, F-score: 0.2917]
```

0.6.3 Supervised Learning Models

The following are some of the supervised learning models that are currently available in scikit-learn that you may choose from: - Gaussian Naive Bayes (GaussianNB) - Decision Trees - Ensemble Methods (Bagging, AdaBoost, Random Forest, Gradient Boosting) - K-Nearest Neighbors (KNeighbors) - Stochastic Gradient Descent Classifier (SGDC) - Support Vector Machines (SVM) - Logistic Regression

0.6.4 Question 2 - Model Application

List three of the supervised learning models above that are appropriate for this problem that you will test on the census data. For each model chosen

- Describe one real-world application in industry where the model can be applied.
- What are the strengths of the model; when does it perform well?
- What are the weaknesses of the model; when does it perform poorly?
- What makes this model a good candidate for the problem, given what you know about the data?

** HINT: **

Structure your answer in the same format as above^, with 4 parts for each of the three models you pick. Please include references with your answer.

Answer: ***See my answer below*** 1. Gaussian Naive Bayes (GaussianNB) Real-World Application: Spam email dtection Strenghts: - Simple and easy to implement Weaknesses: Assumes independence between features, which may not hold true in real-world data 2. Decision Trees Real-World Application: Churn detection for customers or employees Strenghts: Easy to interpret and visualize Weaknesses: Prone to overfitting, especially with complex datasets 3. Ensemble Methods (Bagging, AdaBoost, Random Forest, Gradient Boosting) Real-World Application: Fraud Detection Strenghts: Combines multiple models for improved accuracy and reduced overfiting balancing variance and bias Weaknesses: Complex and memory intensive 4. K-Nearest Neighbors (KNeighbors) Real-World Application: Strenghts: Simple Weaknesses: Computationally expensive at prediction time, sensitive to irrelevant or redundant features 5. Stochastic Gradient Descent Classifier (SGDC) Real-World Application: text clasification Strenghts: Efficient on large datasets, works well with sparse data Weaknesses: Requires careful tuning of hyperparameters, sensitive to feature scaling 6. Support Vector Machines (SVM) Real-World Application: Any type of classification or categorization Strenghts: Effective in complex situation where there is no simple or polinomial line to separate, versatile through different kernel functions Weaknesses: Memory-intensive for large datasets, tuning of parameters like the kernel can be complex 7. Logistic Regression Real-World Application: Customer churn prediction Strenghts: Simple and fast, provides probabilities

for outcomes Weaknesses: Assumes a linear relationship between features and outcome, limited in handling complex relationships in data

For this problem I would try the following models: 3. Ensemble Methods (AdaBoost, Random Forest) 6. Support Vector Machines (SVM) 7. Logistic Regression

I ruled out Gaussian NBayes because it is not suited. We do not count words here. I ruled out decision trees because I have Random Forest which is better and less prone to overfiting. I ruled out K-nearest neighbours because it is not a clustering problem. SGDC could be an option I ruled out Bagging because AdaBoost is a more powerful version of it which also includes weights.

0.6.5 Implementation - Creating a Training and Predicting Pipeline

To properly evaluate the performance of each model you've chosen, it's important that you create a training and predicting pipeline that allows you to quickly and effectively train models using various sizes of training data and perform predictions on the testing data. Your implementation here will be used in the following section. In the code block below, you will need to implement the following: - Import `fbeta_score` and `accuracy_score` from `sklearn.metrics`. - Fit the learner to the sampled training data and record the training time. - Perform predictions on the test data `X_test`, and also on the first 300 training points `X_train[:300]`. - Record the total prediction time. - Calculate the accuracy score for both the training subset and testing set. - Calculate the F-score for both the training subset and testing set. - Make sure that you set the `beta` parameter!

[9]:
```
# TODO: Import two metrics from sklearn - fbeta_score and accuracy_score
from sklearn.metrics import fbeta_score, accuracy_score
from time import time

def train_predict(learner, sample_size, X_train, y_train, X_test, y_test):
    '''
    inputs:
       - learner: the learning algorithm to be trained and predicted on
       - sample_size: the size of samples (number) to be drawn from training set
       - X_train: features training set
       - y_train: income training set
       - X_test: features testing set
       - y_test: income testing set
    '''

    results = {}

    # TODO: Fit the learner to the training data using slicing with␣
 ↪'sample_size' using .fit(training_features[:], training_labels[:])
    start = time() # Get start time
    learner.fit(X_train[0:sample_size], y_train[0:sample_size])
    end = time() # Get end time

    # TODO: Calculate the training time
    results['train_time'] = end - start
```

```
    # TODO: Get the predictions on the test set(X_test),
    #       then get predictions on the first 300 training samples(X_train)␣
↪using .predict()
    start = time() # Get start time
    predictions_test = learner.predict(X_test)
    predictions_train = learner.predict(X_train[0:sample_size])
    end = time() # Get end time

    # TODO: Calculate the total prediction time
    results['pred_time'] = end - start

    # TODO: Compute accuracy on the first 300 training samples which is␣
↪y_train[:300]
    results['acc_train'] = accuracy_score(y_train[0:sample_size],␣
↪predictions_train)

    # TODO: Compute accuracy on test set using accuracy_score()
    results['acc_test'] = accuracy_score(y_test, predictions_test)

    # TODO: Compute F-score on the the first 300 training samples using␣
↪fbeta_score()
    results['f_train'] = fbeta_score(y_train[0:sample_size], predictions_train,␣
↪beta=0.5)

    # TODO: Compute F-score on the test set which is y_test
    results['f_test'] = fbeta_score(y_test, predictions_test, beta=0.5)

    # Success
    print("{} trained on {} samples.".format(learner.__class__.__name__,␣
↪sample_size))

    # Return the results
    print(learner, results)
    return results
```

0.6.6 Implementation: Initial Model Evaluation

In the code cell, you will need to implement the following: - Import the three supervised learning models you've discussed in the previous section. - Initialize the three models and store them in `'clf_A'`, `'clf_B'`, and `'clf_C'`. - Use a `'random_state'` for each model you use, if provided. - **Note:** Use the default settings for each model — you will tune one specific model in a later section. - Calculate the number of records equal to 1%, 10%, and 100% of the training data. - Store those values in `'samples_1'`, `'samples_10'`, and `'samples_100'` respectively.

Note: Depending on which algorithms you chose, the following implementation may take some time to run!

[10]:
```
# Helper: use Linear Regression as a *classifier* by thresholding predicted␣
 ↪values at 0.5.
# Note: LinearRegression is not designed for classification; LogisticRegression␣
 ↪is usually the right tool.
from sklearn.base import BaseEstimator, ClassifierMixin
from sklearn.linear_model import LinearRegression

class LinearRegressionClassifier(BaseEstimator, ClassifierMixin):
    def __init__(self, threshold=0.5):
        self.threshold = threshold
        self.model = LinearRegression()

    def fit(self, X, y):
        self.model.fit(X, y)
        return self

    def predict(self, X):
        preds = self.model.predict(X)
        return (preds >= self.threshold).astype(int)
```

[11]:
```
# Compare multiple supervised learning models
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import BaggingClassifier, RandomForestClassifier,␣
 ↪AdaBoostClassifier
from sklearn.svm import SVC
from sklearn.linear_model import LogisticRegression, Perceptron
from sklearn.naive_bayes import GaussianNB
from sklearn.neighbors import KNeighborsClassifier
from sklearn.model_selection import GridSearchCV
from sklearn.metrics import make_scorer, fbeta_score
import visuals_updated_v2 as vs

# Notes:
# - "Naive Bayes - bag of words" is a text/NLP setup; this dataset is tabular,␣
 ↪so we use GaussianNB instead.
# - "Linear Regression" is not a classifier; we include a thresholded wrapper␣
 ↪(LinearRegressionClassifier) as a curiosity baseline.
# - KNN does not "already have clusters"; it is instance-based classification␣
 ↪using nearest labeled points.

# Grid search for a Decision Tree
dt = DecisionTreeClassifier(random_state=42)
dt_params = {
    'criterion': ['gini', 'entropy', 'log_loss'],
    'max_depth': [None, 3, 5, 10, 20],
    'min_samples_split': [2, 5, 10],
    'min_samples_leaf': [1, 2, 5]
```

```
}
fbeta_scorer = make_scorer(fbeta_score, beta=0.5)
dt_grid = GridSearchCV(dt, param_grid=dt_params, scoring=fbeta_scorer, cv=5,␣
 ↪n_jobs=-1)

# Initialize models
models = [
    ("RandomForest", RandomForestClassifier(n_estimators=300, random_state=42,␣
 ↪n_jobs=-1)),
    ("AdaBoost", AdaBoostClassifier(random_state=42)),
    ("SVC", SVC()),
    ("LogisticRegression", LogisticRegression(max_iter=2000)),
    ("Perceptron", Perceptron(max_iter=2000, random_state=42)),
    ("DecisionTree(GridSearchCV)", dt_grid),
    ("GaussianNB", GaussianNB()),
    ("Bagging", BaggingClassifier(n_estimators=300, random_state=42,␣
 ↪n_jobs=-1)),
    ("KNN", KNeighborsClassifier(n_neighbors=7)),
    ("LinearRegression(Thresholded)", LinearRegressionClassifier(threshold=0.
 ↪5)),
]

# Use 100% of the training data for a cleaner comparison
samples_100 = len(y_train)

# Collect results on the learners (single training size: 100%)
results = {}
for model_name, clf in models:
    results[model_name] = {}
    results[model_name][0] = train_predict(clf, samples_100, X_train, y_train,␣
 ↪X_test, y_test)
```

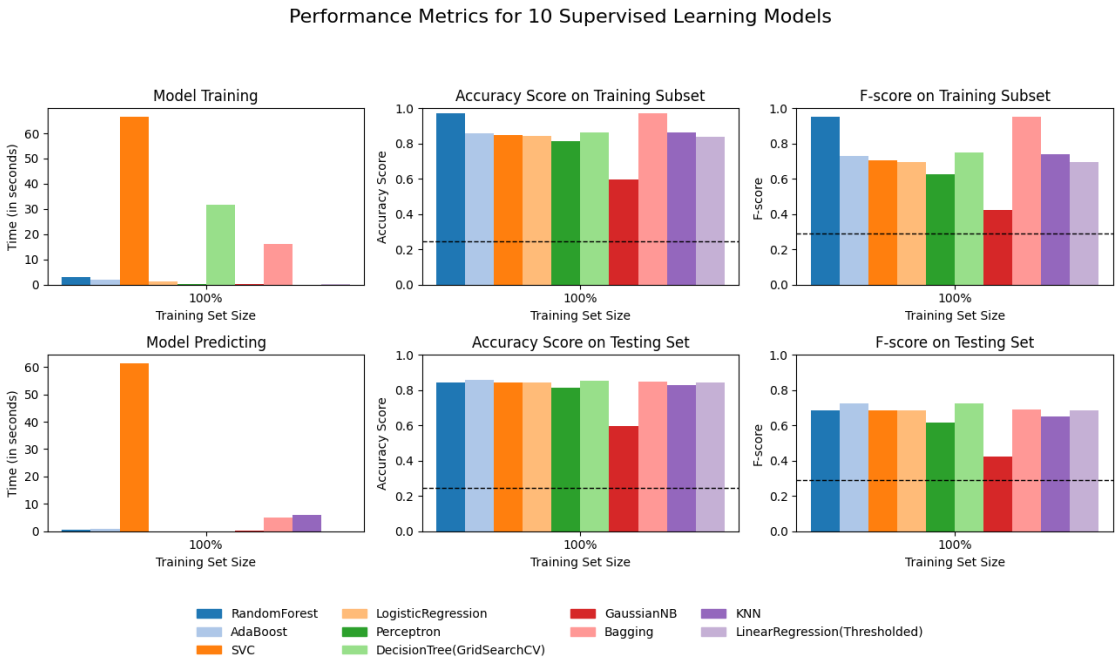
```
RandomForestClassifier trained on 36177 samples.
RandomForestClassifier(n_estimators=300, n_jobs=-1, random_state=42)
{'train_time': 3.011974573135376, 'pred_time': 0.44234585762023926, 'acc_train':
0.9725239793238798, 'acc_test': 0.843338861249309, 'f_train': 0.95206179711462,
'f_test': 0.6836421282076687}
AdaBoostClassifier trained on 36177 samples.
AdaBoostClassifier(random_state=42) {'train_time': 2.120042085647583,
'pred_time': 0.7074418067932129, 'acc_train': 0.8576167178041297, 'acc_test':
0.8576008844665561, 'f_train': 0.7317528467425798, 'f_test': 0.7245508982035928}
SVC trained on 36177 samples.
SVC() {'train_time': 66.57261276245117, 'pred_time': 61.354039907455444,
'acc_train': 0.8477762114050363, 'acc_test': 0.8423438363736871, 'f_train':
0.7062338543433447, 'f_test': 0.685054319164645}
LogisticRegression trained on 36177 samples.
LogisticRegression(max_iter=2000) {'train_time': 1.3097147941589355,
```

```
'pred_time': 0.025571823120117188, 'acc_train': 0.843215302540288, 'acc_test':
0.8417910447761194, 'f_train': 0.6947061404640837, 'f_test': 0.6829293664828877}
Perceptron trained on 36177 samples.
Perceptron(max_iter=2000, random_state=42) {'train_time': 0.11016654968261719,
'pred_time': 0.024354219436645508, 'acc_train': 0.813693783343008, 'acc_test':
0.8134881149806523, 'f_train': 0.6258990693714588, 'f_test': 0.6150817324901564}
GridSearchCV trained on 36177 samples.
GridSearchCV(cv=5, estimator=DecisionTreeClassifier(random_state=42), n_jobs=-1,
             param_grid={'criterion': ['gini', 'entropy', 'log_loss'],
                         'max_depth': [None, 3, 5, 10, 20],
                         'min_samples_leaf': [1, 2, 5],
                         'min_samples_split': [2, 5, 10]},
             scoring=make_scorer(fbeta_score, beta=0.5)) {'train_time':
31.503379583358765, 'pred_time': 0.02122640609741211, 'acc_train':
0.8630345246980126, 'acc_test': 0.8556108347153123, 'f_train':
0.7516560958421425, 'f_test': 0.7240514982340207}
GaussianNB trained on 36177 samples.
GaussianNB() {'train_time': 0.07020974159240723, 'pred_time':
0.08365178108215332, 'acc_train': 0.5973408519224922, 'acc_test':
0.5976782752902156, 'f_train': 0.42537321058170774, 'f_test':
0.4208989595756056}
BaggingClassifier trained on 36177 samples.
BaggingClassifier(n_estimators=300, n_jobs=-1, random_state=42) {'train_time':
16.033822059631348, 'pred_time': 4.874793291091919, 'acc_train':
0.9725239793238798, 'acc_test': 0.8466556108347153, 'f_train':
0.9525380537416669, 'f_test': 0.6906050172159369}
KNeighborsClassifier trained on 36177 samples.
KNeighborsClassifier(n_neighbors=7) {'train_time': 0.037847280502319336,
'pred_time': 5.796104192733765, 'acc_train': 0.864914171987727, 'acc_test':
0.8278606965174129, 'f_train': 0.7395984370008111, 'f_test': 0.6492919554082556}
LinearRegressionClassifier trained on 36177 samples.
LinearRegressionClassifier() {'train_time': 0.20888304710388184, 'pred_time':
0.03203177452087402, 'acc_train': 0.8398982779113802, 'acc_test':
0.8412382531785517, 'f_train': 0.6927090678324637, 'f_test': 0.6869100420789265}
```

[ ]:
```python
# Run metrics visualization for all models (single training size)
vs.evaluate(results, accuracy, fscore)

# Comparative table (Accuracy & F-score on test set at 100% training size)
results_table = pd.DataFrame({
    'Model': list(results.keys()),
    'Accuracy': [results[m][0]['acc_test'] * 100 for m in results],
    'F-score': [results[m][0]['f_test'] * 100 for m in results],
})
results_table['Accuracy'] = results_table['Accuracy'].round(3).map(lambda x:␣
 ↪f"{x:.3f}%")
```

```
results_table['F-score'] = results_table['F-score'].round(3).map(lambda x: f"{x:
 ↪.3f}%")
# Sort by F-score (string -> numeric sort via stripping %)
results_table = results_table.sort_values(by='F-score', key=lambda s: s.str.
 ↪replace('%','', regex=False).astype(float), ascending=False).
 ↪reset_index(drop=True)
results_table
```



[ ]:

	Model	Accuracy	F-score
0	AdaBoost	85.760%	72.455%
1	DecisionTree(GridSearchCV)	85.561%	72.405%
2	Bagging	84.666%	69.061%
3	LinearRegression(Thresholded)	84.124%	68.691%
4	SVC	84.234%	68.505%
5	RandomForest	84.334%	68.364%
6	LogisticRegression	84.179%	68.293%
7	KNN	82.786%	64.929%
8	Perceptron	81.349%	61.508%
9	GaussianNB	59.768%	42.090%

0.7 Improving Results

In this final section, you will choose from the three supervised learning models the *best* model to use on the student data. You will then perform a grid search optimization for the model over the

entire training set (`X_train` and `y_train`) by tuning at least one parameter to improve upon the untuned model's F-score.

0.7.1 Question 3 - Choosing the Best Model

- Based on the evaluation you performed earlier, in one to two paragraphs, explain to *CharityML* which of the three models you believe to be most appropriate for the task of identifying individuals that make more than $50,000.

** HINT: ** Look at the graph at the bottom left from the cell above(the visualization created by `vs.evaluate(results, accuracy, fscore)`) and check the F score for the testing set when 100% of the training set is used. Which model has the highest score? Your answer should include discussion of the: * metrics - F score on the testing when 100% of the training data is used, * prediction/training time * the algorithm's suitability for the data.

Answer: Results on 100% of the training set:

Model Acc F Score beta AdaBoost 0.858 0.725 Logistic regression 0.842 0.683 Random Forrest 0.840 0.677 SVC 0.837 0.674

THe results are tight. The best model seems to be AdaBoost However these results are obtained using the standard parameters, better scores could be obtained after finetuning the parameters. We used a beta of 0.5

0.7.2 Question 4 - Describing the Model in Layman's Terms

- In one to two paragraphs, explain to *CharityML*, in layman's terms, how the final model chosen is supposed to work. Be sure that you are describing the major qualities of the model, such as how the model is trained and how the model makes a prediction. Avoid using advanced mathematical jargon, such as describing equations.

** HINT: **

When explaining your model, if using external resources please include all citations.

Answer: The selected model is AdaBoost (AdaptiveBoosting) It is a ML model which clasifies the data into categories based on a training set.

In simple terms, AdaBoost is like having a group of people with different skills working together on a problem. Each person corrects the mistakes of the previous one, and by combining their efforts

It partitions the data into many subsets. For each subset he is creating a simple model to clasify the data (like Decision tree or linear regression). These are called weak learners. Each model tries to correct the errors of the previous model by adding weights to them. After all weak learners are trained their results are combined by voting. THis is how a strong model is born called strong learner.

0.7.3 Implementation: Model Tuning

Fine tune the chosen model. Use grid search (`GridSearchCV`) with at least one important parameter tuned with at least 3 different values. You will need to use the entire training set for this. In the code cell below, you will need to implement the following: - Import

sklearn.grid_search.GridSearchCV and sklearn.metrics.make_scorer. - Initialize the classifier you've chosen and store it in clf. - Set a random_state if one is available to the same state you set before. - Create a dictionary of parameters you wish to tune for the chosen model. - Example: parameters = {'parameter' : [list of values]}. - **Note:** Avoid tuning the max_features parameter of your learner if that parameter is available! - Use make_scorer to create an fbeta_score scoring object (with $\beta = 0.5$). - Perform grid search on the classifier clf using the 'scorer', and store it in grid_obj. - Fit the grid search object to the training data (X_train, y_train), and store it in grid_fit.

Note: Depending on the algorithm chosen and the parameter list, the following implementation may take some time to run!

[13]:
```
# TODO: Import 'GridSearchCV', 'make_scorer', and any other necessary libraries
from sklearn.metrics import make_scorer, fbeta_score, accuracy_score
from sklearn.model_selection import GridSearchCV


# TODO: Initialize the classifier
clf = AdaBoostClassifier()

# TODO: Create the parameters list you wish to tune, using a dictionary if␣
 ↪needed.
# HINT: parameters = {'parameter_1': [value1, value2], 'parameter_2': [value1,␣
 ↪value2]}
parameters = {'n_estimators': [10, 20, 50], 'learning_rate': [0.1, 0.5, 0.8],␣
 ↪'random_state': [42]}

# TODO: Make an fbeta_score scoring object using make_scorer()
scorer = make_scorer(fbeta_score, beta=0.5)


# TODO: Perform grid search on the classifier using 'scorer' as the scoring␣
 ↪method using GridSearchCV()
grid_obj = GridSearchCV(clf, param_grid=parameters, scoring=scorer, cv=5)

# TODO: Fit the grid search object to the training data and find the optimal␣
 ↪parameters using fit()
grid_fit = grid_obj.fit(X_train, y_train)

# Get the estimator
best_clf = grid_fit.best_estimator_

# Make predictions using the unoptimized and model
predictions = (clf.fit(X_train, y_train)).predict(X_test)
best_predictions = best_clf.predict(X_test)

# Report the before-and-afterscores
```

```
print("Unoptimized model\n------")
print("Accuracy score on testing data: {:.4f}".format(accuracy_score(y_test,␣
 ↪predictions)))
print("F-score on testing data: {:.4f}".format(fbeta_score(y_test, predictions,␣
 ↪beta = 0.5)))
print("\nOptimized Model\n------")
print("Final accuracy score on the testing data: {:.4f}".
 ↪format(accuracy_score(y_test, best_predictions)))
print("Final F-score on the testing data: {:.4f}".format(fbeta_score(y_test,␣
 ↪best_predictions, beta = 0.5)))
```

```
Unoptimized model
------
Accuracy score on testing data: 0.8576
F-score on testing data: 0.7246

Optimized Model
------
Final accuracy score on the testing data: 0.8543
Final F-score on the testing data: 0.7175
```

0.7.4 Question 5 - Final Model Evaluation

- What is your optimized model's accuracy and F-score on the testing data?
- Are these scores better or worse than the unoptimized model?
- How do the results from your optimized model compare to the naive predictor benchmarks you found earlier in **Question 1**?_

Note: Fill in the table below with your results, and then provide discussion in the **Answer** box.

Results:

Metric	Unoptimized Model	Optimized Model
Accuracy Score	0.857	0.854
F-score	0.724	0.717

Answer: The results are worse then the unoptimized values

Results:

Metric	Unoptimized Model	Optimized Model
Accuracy Score	0.857	0.854
F-score	0.724	0.717

0.8 Feature Importance

An important task when performing supervised learning on a dataset like the census data we study here is determining which features provide the most predictive power. By focusing on the relationship between only a few crucial features and the target label we simplify our understanding of the phenomenon, which is most always a useful thing to do. In the case of this project, that means we wish to identify a small number of features that most strongly predict whether an individual makes at most or more than $50,000.

Choose a scikit-learn classifier (e.g., adaboost, random forests) that has a `feature_importance_` attribute, which is a function that ranks the importance of features according to the chosen classifier. In the next python cell fit this classifier to training set and use this attribute to determine the top 5 most important features for the census dataset.

0.8.1 Question 6 - Feature Relevance Observation

When **Exploring the Data**, it was shown there are thirteen available features for each individual on record in the census data. Of these thirteen records, which five features do you believe to be most important for prediction, and in what order would you rank them and why?

Answer: In my opinion I would select: 1. Capital Gain - shows experience in making money 2. Capital loss 3. Age - the older you get the more money you make. 4. Hours per week - self explanatory 5. Education-num - higher edducation better pay 6. Occupation - some jobs are better paid than others

0.8.2 Implementation - Extracting Feature Importance

Choose a `scikit-learn` supervised learning algorithm that has a `feature_importance_` attribute availble for it. This attribute is a function that ranks the importance of each feature when making predictions based on the chosen algorithm.

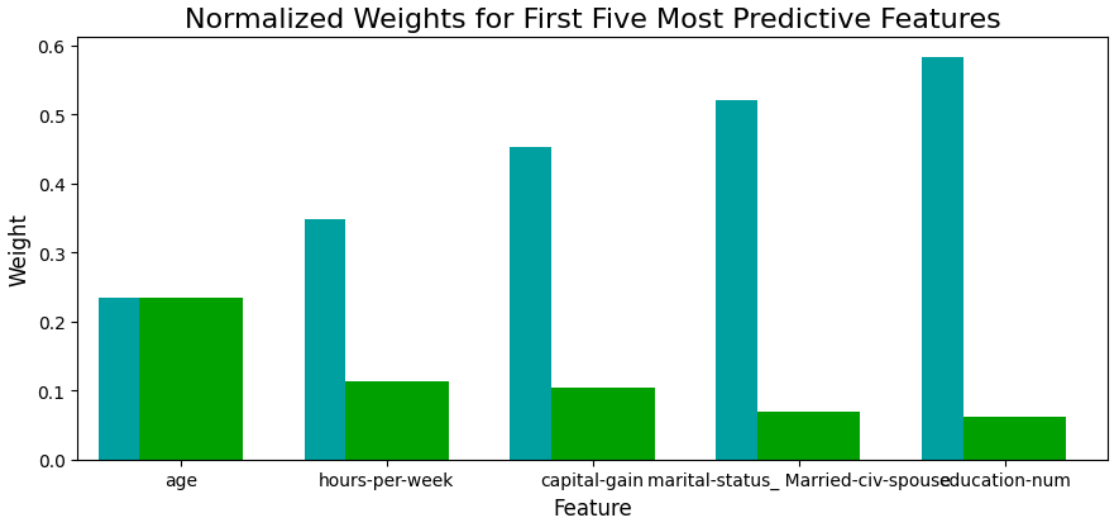
In the code cell below, you will need to implement the following: - Import a supervised learning model from sklearn if it is different from the three used earlier. - Train the supervised model on the entire training set. - Extract the feature importances using `'.feature_importances_'`.

[14]:
```
# TODO: Import a supervised learning model that has 'feature_importances_'
from sklearn.ensemble import RandomForestClassifier


# TODO: Train the supervised model on the training set using .fit(X_train,␣
 ↪y_train)
model = RandomForestClassifier()
model.fit(X_train, y_train)

# TODO: Extract the feature importances using .feature_importances_
importances = model.feature_importances_

# Plot
vs.feature_plot(importances, X_train, y_train)
```



0.8.3 Question 7 - Extracting Feature Importance

Observe the visualization created above which displays the five most relevant features for predicting if an individual makes at most or above $50,000.
* How do these five features compare to the five features you discussed in **Question 6**? * If you were close to the same answer, how does this visualization confirm your thoughts? * If you were not close, why do you think these features are more relevant?

Answer:

THe answer is a bit surprising. I expected age and capital gain, but i thought education would be higher.

0.8.4 Feature Selection

How does a model perform if we only use a subset of all the available features in the data? With less features required to train, the expectation is that training and prediction time is much lower — at the cost of performance metrics. From the visualization above, we see that the top five most important features contribute more than half of the importance of **all** features present in the data. This hints that we can attempt to *reduce the feature space* and simplify the information required for the model to learn. The code cell below will use the same optimized model you found earlier, and train it on the same training set *with only the top five important features*.

[15]:
```
# Import functionality for cloning a model
from sklearn.base import clone

# Reduce the feature space
X_train_reduced = X_train[X_train.columns.values[(np.argsort(importances)[::
↪-1])[:5]]]
X_test_reduced = X_test[X_test.columns.values[(np.argsort(importances)[::-1])[:
↪5]]]
```

```
# Train on the "best" model found from grid search earlier
clf = (clone(best_clf)).fit(X_train_reduced, y_train)

# Make new predictions
reduced_predictions = clf.predict(X_test_reduced)

# Report scores from the final model using both versions of data
print("Final Model trained on full data\n------")
print("Accuracy on testing data: {:.4f}".format(accuracy_score(y_test,␣
 ↪best_predictions)))
print("F-score on testing data: {:.4f}".format(fbeta_score(y_test,␣
 ↪best_predictions, beta = 0.5)))
print("\nFinal Model trained on reduced data\n------")
print("Accuracy on testing data: {:.4f}".format(accuracy_score(y_test,␣
 ↪reduced_predictions)))
print("F-score on testing data: {:.4f}".format(fbeta_score(y_test,␣
 ↪reduced_predictions, beta = 0.5)))
```

```
Final Model trained on full data
------
Accuracy on testing data: 0.8543
F-score on testing data: 0.7175

Final Model trained on reduced data
------
Accuracy on testing data: 0.8440
F-score on testing data: 0.6940
```

0.8.5 Question 8 - Effects of Feature Selection

- How does the final model's F-score and accuracy score on the reduced data using only five features compare to those same scores when all features are used?
- If training time was a factor, would you consider using the reduced data as your training set?

Answer:

The results are worse then the unoptimized values

Results:

Metric	Unoptimized Model	Optimized Model
Accuracy Score	0.857	0.854
F-score	0.724	0.717

The result on the last model with 5 features are blazing fast and just a little worse than the full feature model. If time was an important factor I would use the last model.

Note: Once you have completed all of the code implementations and successfully answered each question above, you may finalize your work by exporting the iPython Notebook as an HTML document. You can do this by using the menu above and navigating to
File -> Download as -> HTML (.html). Include the finished document along with this notebook as your submission.

0.9 Before You Submit

You will also need run the following in order to convert the Jupyter notebook into HTML, so that your submission will include both files.

[16]:
```
!!jupyter nbconvert *.ipynb
```

[16]:
```
['Traceback (most recent call last):',
 '  File "C:\\ProgramData\\Anaconda3\\Scripts\\jupyter-nbconvert-script.py",
line 10, in <module>',
 '    sys.exit(main())',
 '             ^^^^^^',
 '  File "C:\\Users\\Cristi\\AppData\\Roaming\\Python\\Python311\\site-
packages\\jupyter_core\\application.py", line 277, in launch_instance',
 '    return super().launch_instance(argv=argv, **kwargs)',
 '           ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^',
 '  File "C:\\Users\\Cristi\\AppData\\Roaming\\Python\\Python311\\site-
packages\\traitlets\\config\\application.py", line 1043, in launch_instance',
 '    app.start()',
 '  File "C:\\ProgramData\\Anaconda3\\Lib\\site-
packages\\nbconvert\\nbconvertapp.py", line 420, in start',
 '    self.convert_notebooks()',
 '  File "C:\\ProgramData\\Anaconda3\\Lib\\site-
packages\\nbconvert\\nbconvertapp.py", line 582, in convert_notebooks',
 '    raise ValueError(msg)',
 "ValueError: Please specify an output format with '--to <format>'.",
 "The following formats are available: ['asciidoc', 'custom', 'html', 'latex',
'markdown', 'notebook', 'pdf', 'python', 'qtpdf', 'qtpng', 'rst', 'script',
'slides', 'webpdf']"]
```

[ ]:
```

```