Machine Learning for Public Policy

Mid-term Assignment

Name: Claire Herdeman

Due: May 8, 5pm

Instructions:

- This is an individual assignment – please do not work in groups.
- This is open-book, open-notes, open-internet but no need for any programming to do any of the work.
- You should show your work instead of just giving me the answer.
- You can spend as much time as you want.
- Please submit the assignment on canvas as a pdf.

(Short answers) [30 pts – 2 points each]

1. You're asked to predict the probability that the unemployment rate will go down next quarter for each of the neighborhoods in Chicago. Which model would you prefer to use?

   A. Logistic Regression
   B. Support Vector Machines

   Why?
   Logistic regression performs better than SVM when there are a large number of features relative to the number of observations. There are only 77 neighborhoods in Chicago, so if we are using a single year of training data our sample size is limited. We could imagine a large number of features to describe a given neighborhood (demographics, distance from downtown or amenities, etc.)

2. Do you have to do anything special with the data for this problem with the model you chose in #1?

   Since this is a regression model, you must be more aware of possible collinearities in your data set. Additionally, this model assumes that that data is linear in the learned parameters.

3. What is the training error (error on the training set) for a 1-NN classifier?

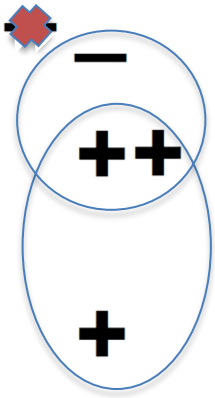
The training error for a 1-NN classifier is 0, because the nearest neighbor will always be itself.

4. What is the Leave-one-out cross validation error for k nearest neighbor on the following data set? List any assumptions you may be making.

   A. For k=1
      0, per above.
   B. For k=3
      0.2 or 20%, see reasoning below.

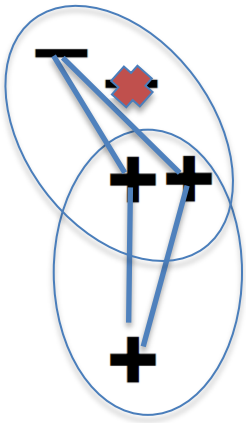
   I'm assuming a Euclidean distance metric, i.e. an "as the crow flies" linear measurement of distance and measurement from the center of the marker.

1.



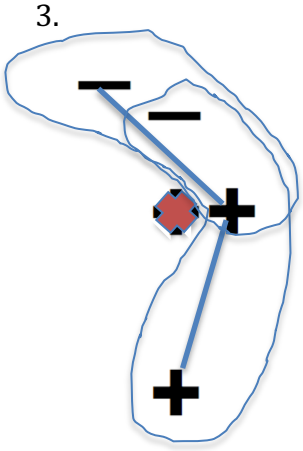
In this case, the bottom + has the 3 nearest neighbors indicated by the oval, and the remaining 3 are circumscribed by the circle. Only the – will be incorrectly classified (as 2/3 of its nearest neighbors are +), so the training error is ¼.

2.



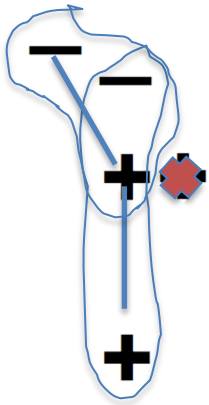
Based on the crude distance calculation shown, the two central + are closer to the – than the bottom plus. Again, the training error will be ¼ as only the – is misclassified.

3.



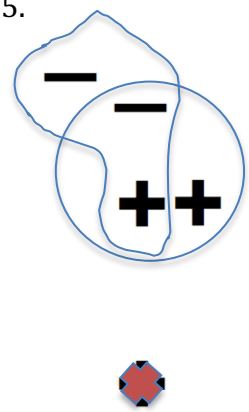
The bottom blob circumscribes the 3NN for the bottom +, while the top blob is the 3NN for the 3 other points. This means that the middle + will be misclassified, leading to training error of ¼.

4.



This results in the same scenario as above with the middle + misclassified, so the training error is ¼.

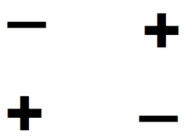
5.



The two +'s will have the 3NN circumscribed by the circle, while the 2 –'s are circumscribed by the amorphous blob. In all of these cases the marker will be correctly classified, so the training error is 0.

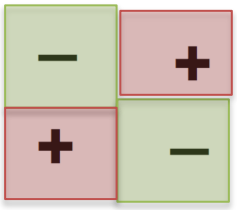
To find the total error we average these values, i.e. (1/4+1/4+1/4+1/4)/5 = 1/5 = 0.2.

5. Which of the following classifiers are appropriate to use for the following data set? Why?



   A. Logistic Regression
   B. Decision Trees
   C. SVMs

   Decision trees. A decision tree will do the best job since it creates partitions that are parallel to the x and y axes, so would effectively be able to draw correct categorization partitions around each point, per the diagram below:



   Neither of the linear models would work well since this dataset is not linearly separable. Any line drawn would misclassify some observations. Logistic regression may produce better output since it would more effectively incorporate all data points into them model.

6. You are being asked to build a model to predict which children in Chicago are at risk of Asthma attacks. You create 1000 features (all continuous) but you find out after exploring the data and talking to public health and medical experts that ~10 of them are useful and relevant for the prediction problem. Which of the classifiers below would you choose? And why?
   a. K-NN
   b. Decision Trees

I would choose a decision tree in this case. Since we know that the ~10 most useful features have been identified, the tree classifier will not be destabilized by irrelevant information. Given a reasonable sample size, the danger of overfitting the tree is not as great as it would be if there were a greater number of features included.

7. Does Boosting give you a linear classifier? Why or why not?

   No, it is very unlikely to be linear unless the data is linearly separable. The final boosting decision boundary will be a piecewise combination of the decision boundaries determined in each step of the boosting process.

8. Can boosting perfectly classify all the training examples for any given data set? Why or why not?

If the data is linearly separable, then a boosting method could perfectly classify all data points. If it is not, however, it will be impossible for all of the training examples to be correctly classified.

9. If you have a data set with 10 variables, each of them binary, with a binary target variable (label). You are asked to build a decision tree. If you didn't use a greedy approach and built all possible trees on this data set (without pruning or limiting the depth), how many trees would you build?

2^(2^10) = 2^1024. This shows us that even with a relatively small number of features, it is an intractably large problem to learn all possible trees and select the best one.

10. You are reading a paper for a new method to identify people at risk of pre-term and adverse births.. The reported accuracy is 89.4% and the precision at the top 10% is 56%. Are those numbers high enough to justify you trying the method in your work (please explain your answer in 1-2 sentences)?
    a. Yes
    b. No
    c. Maybe

No, there is not enough information here. For example, if 89.4% of births are non-adverse and the remaining are adverse, a model with 89.4% accuracy may be classifying all births as non-adverse. What seems like a high accuracy score is a

result of the data itself rather than any additional information gained from the model. Neither of these metrics gives us insight into the false negative rate in the model, which is extremely important to understand in the case of adverse births. Precision of 56% at the top 10% does not seem particularly high; if there are 100 births in this decile, 44 of them will be false positives. This seems like a high proportion of false alarms when we do not know how many false negatives we are missing.

11. A Random Forest will always perform better than a decision tree on the same data set.

A. True

B. False

True. While I'm sure that counterexamples could be created, a random forest will almost always perform better than a single decision tree on the same data set. This is due to its greater robustness and stability.

12-15. You need to build a model to predict re-entry into a social service program. A colleague suggests building a separate model for males and females while another colleague insists you just need to build one combined model.

12. When will separate models be more appropriate?

Separate models may be more appropriate if you believe that the underlying re-entry rate will be different for men and women, and if you will intervene differently within the male and female population based on this knowledge.

13. When will a combined model be more appropriate?

When the converse is true, when you believe that the re-entry rate for both males and females is the same, or if you have a single intervention that will be applied to those most likely to re-enter regardless of gender.

14. What are the pros and cons of each approach?

The gender-specific approach allows for true variation in the rate of re-entry and an equitable split of intervention resources to address the differing issues of the genders. However, this additional expenditure or effort may or may not be worth the cost if the goal is to intervene for those most likely to re-enter, regardless of gender. The single-model approach will more efficiently find those more likely to re-enter, but if either gender is much more likely to re-enter, the model may overpredict re-entry for that group.

15. What is your opinion on how to proceed?

As is usually the case, it depends. In this case, the desired mode of intervention, desired outcome of intervention, and underlying population characteristics will all play a role. More context on the specifics of the problem is needed to make a determination.

Section B [55 pts]

1. Decision Trees [12 pts]

Temperature	HomeInsulation	HomeSize	EnergyConsumption
Hot	Poor	Small	Low
Mild	Poor	Medium	High
Cool	Excellent	Large	Low
Hot	Excellent	Large	High
Hot	Excellent	Medium	Low
Mild	Poor	Small	High
Cool	Poor	Small	High
Cool	Excellent	Medium	Low
Cool	Excellent	Medium	High
Cool	Poor	Medium	High

A. What will be the random baseline accuracy for this data set?

In this sample, p(low) = 0.4 and p(high) = 0.6. Random baseline accuracy will be calculated according to the formula acc = p(low)^2 + p(high)^2 = 0.4^2+0.6^2 = 0.52.

B. Calculate the entropy for the target variable, EnergyConsumption

Note: calculations were done in excel, also attached to submission

Entropy = $\sum_i -p_i log_2 (p_i)$

In this case that means entropy of Energy Consumption = -0.4$log_2(0.4)$ – 0.6$log_2(0.6)$ = 0.971

C. Now calculate the Information Gain if you do a split on the feature "Home Insulation".

IG = entropy(parent) – [weighted avg]entropy(children)

Entropy(parent) = entropy(EC) = 0.971

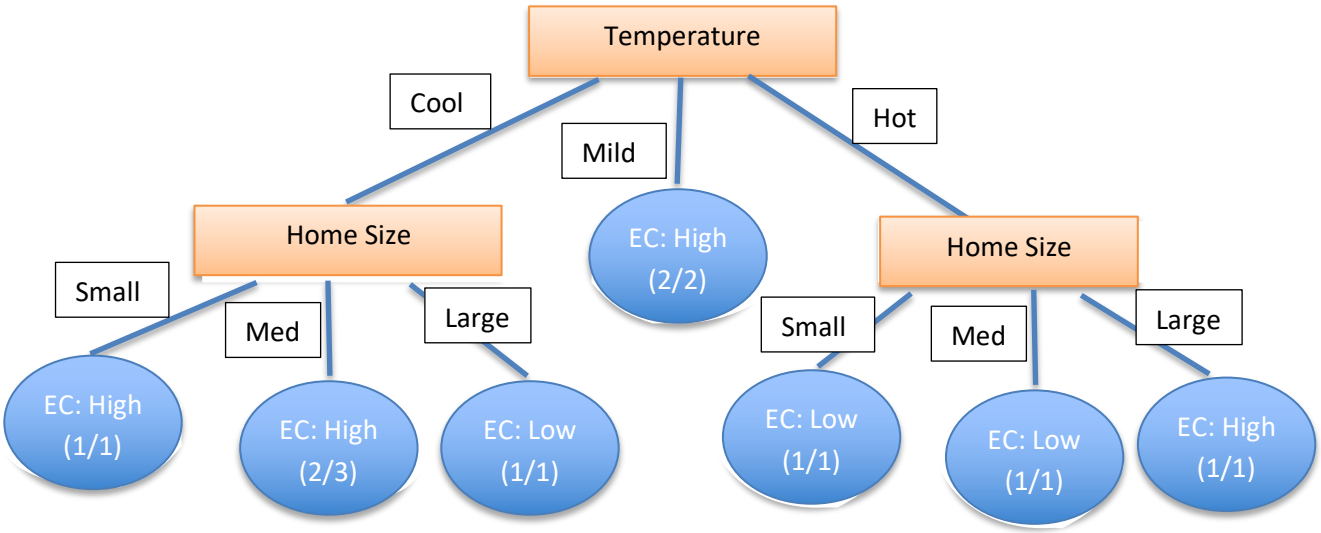
When we split on Home Insulation, we get the EC observations for poor insulation are (LHHHH) and the observations for excellent insulation are (LLLHH). This the weighted entropy calculation is:

$0.5(-0.2log_2(0.2) -0.8log_2(0.8)) + 0.5(-0.4log_2(0.4) -0.6log_2(0.6))$

= 0.846

Information gain = 0.971-0.846 = 0.125

D. Using the data above, construct a two-level decision tree that can be used to predict Energy Consumption. Don't worry about overfitting or pruning. You can use a simple algorithm such as ID3 (using information gain as the splitting criterion).



Entropy/information gain was calculated in excel. Overall entropy for each feature was temperature (0.761), home size (0.961), and insulation (0.846). Temperature had the lowest entropy/highest information gain so was chosen as the first split. Mild temperature already produced a pure result so no additional split was added. For each of cool and hot, entropy/information gain of home size and insulation were calculated. In both instances home size had lower entropy/higher IG so was chosen as the split. All leaf notes except for Cool temperature > Medium home size produce a pure result. This likely means that the model is overfit, but we are not concerned for the purposes of this exercise.

2. Evaluation 1 [12 pts]

The table below shows the predictions of two classifiers, SVM and Logistic Regression for 10 examples. The classifiers are predicting the probability that the Label is 1.

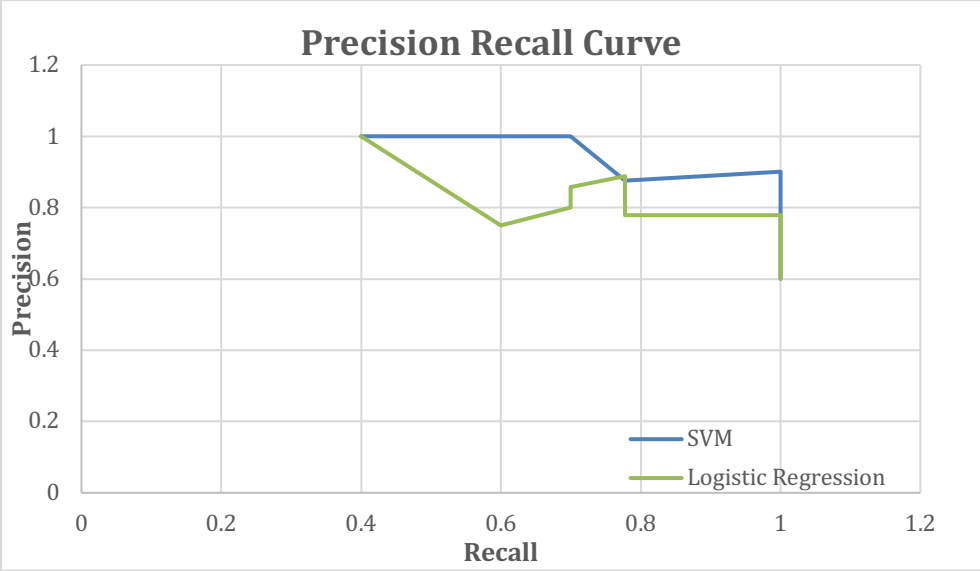
ID	Probability (assigned by SVM)	Probability (assigned by Logistic Regression)	True Label
1	0.98	0.85	1
2	0.2	0.3	0
3	0.1	0.22	0
4	0.99	0.9	1
5	0.55	0.4	0
6	0.05	0.2	0
7	0.4	0.1	1
8	0.35	0.35	0
9	0.65	0.81	0
10	0.75	0.5	1

A. What is the accuracy of the SVM on this set? You will need to make some assumptions here. Be very explicit about your assumptions
Assume that the classification threshold is 0.6, i.e. a probability >0.6 is classified as a 1.
Then the accuracy is 80% according to the confusion matrix below:

		True Label	
		1	0
Predicted	1	3	1
	0	1	5

Acc = (TP + TN) / all observation = 8/10 = 0.8

B. Plot the precision recall curves for both classifiers based on these predictions.



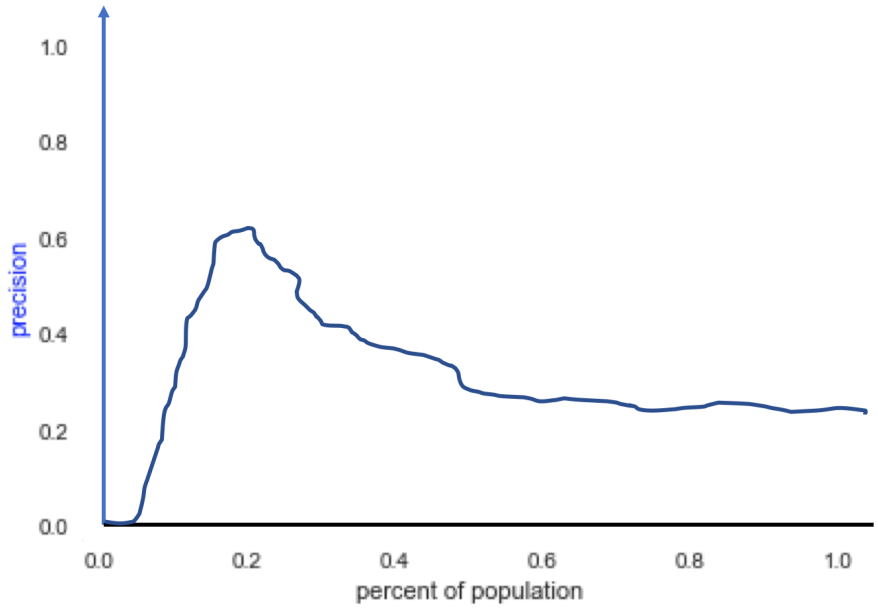
Note: I calculated these curves using increments of 0.1, smaller increments would produce more fine-grained curves.

C. Which classifier is better? (again, list the assumptions you're making)

SVM is better for the entire area in which the curve is above the logistic regression curve. That means that for almost all thresholds, the SVM model has both higher precision and higher recall than the logistic regression model.

3. Evaluation 2 [12 pts]

You just finished building and evaluating a model and got the following precision graph. You might notice that the precision is very low in the beginning of that graph.



A. How would you explain what's happening at the beginning of the graph to someone who's not a machine learning expert?

Precision is telling us how often we are correct out of all the times that we guess that a piece of data belongs in the "positive" class.

Very simplistically, let's say that our goal is to classify all children into whether or not they get Type I diabetes. When our percent of population or "threshold" is 0, our model guesses that all children get Type I diabetes. As you can imagine, almost all of these "positive" predictions will be wrong. Therefore, our precision will be close to 0.

B. What could be the reason for that behavior?

One possible explanation for this behavior is that we have a large number of observations in our data set (i.e., all children), but a very small number are part of the positive class (i.e., are diagnosed with type I diabetes). When we say that everything is true, i.e. all children have type I diabetes, we are almost always wrong. This means that our precision will be almost 0. As we begin to discriminate a bit more, we capture more of the class correctly which improves our precision.

C. What would you do to improve the performance of the classifier at the top 5%?

We could consider all observations in the top 5% separately from the other 95%. Then we could redraw the precision curve based only on the top 5% and find the threshold that performs best on those in the highest risk group.

4. Evaluation 3 [10 pts]

You have trained three types of models on a training set and validated the results on a hold-out test set on a variety of metrics. The table below shows results from your trials.

Model Type	parameters	baseline	Precision at 5%	Precision at 10%	Precision at 20%	AUC ROC
Decision Trees	'max_depth': 1	0.34	0.36	0.32	0.40	0.50
Decision Trees	'max_depth': 20	0.34	0.27	0.25	0.27	0.45
Decision Trees	'max_depth': 100	0.34	0.30	0.33	0.31	0.49
Logistic Regression	{'penalty': 'l1', 'C': 0.001}	0.34	0.66	0.60	0.42	0.53
Logistic Regression	{'penalty': 'l2', 'C': 0.001}	0.34	0.73	0.53	0.45	0.57
Logistic Regression	{'penalty': 'l1', 'C': 0.1}	0.34	0.82	0.64	0.52	0.62
Logistic Regression	{'penalty': 'l2', 'C': 0.1}	0.34	0.82	0.65	0.53	0.63
Logistic Regression	{'penalty': 'l1', 'C': 1}	0.34	0.68	0.57	0.49	0.62
Logistic Regression	{'penalty': 'l2', 'C': 1}	0.34	0.77	0.68	0.54	0.62
Logistic Regression	{'penalty': 'l1', 'C': 10}	0.34	0.70	0.60	0.51	0.62
Logistic Regression	{'penalty': 'l2', 'C': 10}	0.34	0.75	0.65	0.51	0.62
Random Forests	{ 'n_estimators': 1000}	0.34	0.59	0.55	0.47	0.61
Random Forests	{'n_estimators': 10000}	0.34	0.57	0.56	0.47	0.61

A. What can you say about the behavior of Logistic Regression as you vary the parameters?

The precision of the logistic regression model varies most in C, which tunes the "strength" of the regularization component. Smaller values of C indicate greater regularization strength. The mid-range values of C (0.1, 1) produce more precise results indicating that moderate regularization strength is optimal. In almost all cases, l2 regularization outperforms l1 regularization. At the top 5%, a C value of 0.1 produces the most precise results regardless of whether the regularization is l1 or l2, while at 10% and 20% a C value of 1 paired with l2 regularization is optimal.

B. Which model would you select to deploy going forward if:

I) your goal was to prioritize 5% of the population to intervene with?

Logistic Regression with C = 0.1 and l2 regularization produces the best results at 5%.

2)the resources available for interventions were yet to be determined?

If the resources were yet to be determined, I would use LR with C = 1 and l2 regularization. Although the precision is slightly lower at 5%, it is higher at both

10% and 20% which means that a greater number of people targeted for intervention would be correct if more resources were available.

5. Communicating your results [9 pts]

You have recently built a model that assigns a risk score to all students beginning 9th grade of not graduating high school within 4 years. You receive a call from the school administrator who asks you "According to your model, Jenny has a risk score of 50 (out of 100), but I know she is a bright student and has done well so far. Why has your model assigned her a score of 50 and not much lower?"

A. How would you explain this to the school administrator? Assume this administrator is a reasonable, intelligent person with extensive school administration experience and little or no background in statistics and machine learning.

Our model takes a number of factors into account, including factors that are outside of her past school performance. A score of 50 is not considered "high-risk" and certainly does not mean that Jenny will drop out, but more students with her background and school performance than you might expect have dropped out in the past. As she progresses through high school, her risk score will likely decrease if she continues to do well, but you may also consider making sure that she has the academic support that she needs and a person that she feels comfortable reaching out to if she begins to struggle for any reason. Again, our goal is to support the success of as many students as possible. The score is one factor among many that you should consider, but we believe that it provides additional information that may not be obviously visible based on what we know about Jenny today.

B. Then suggest a different way that the administrator can confirm the accuracy of the predictive model you created.

Since this project has been ongoing for over 4 years (making a leap here without this exact information), consider going back to some of our earlier predictions and look at how students with a score of 50 progressed over time. I'd also take a look at students with high and low scores to understand if those groups were predicted correctly in the past. No prediction model will be perfect but we do believe that these scores are informative.

Section C: Solving a New Problem [15 pts]

- Please do not use the internet or books for this question.

A critical part of most machine learning problems is integrating data from multiple sources about the same people, places, or businesses. For example, if you are working on predicting the risk of an individual going back to jail in order to inform preventative interventions, you might get data about that individual from the Department of Corrections, Department of Mental Health, Emergency Medical Services, and Homeless Shelters. Your first task is often to integrate that data and link records about the same people. This is known as record linkage (or matching) and is often done through exact matching or through "fuzzy" matching rules.

Now that you know how to do machine learning, your task is to come up with a machine learning solution for this problem of linking records that belong to the same person across different data sources.

You can assume that all data sources have some columns/fields in common, let's say first name, last name, date of birth, address, gender, and race.

A. How would you formulate this as a machine learning problem? (is it supervised learning or unsupervised learning? If it's the former, what's the label? What is each row in your training data? How would you get the training data?

I would build on the exact and fuzzy matching record linking methodologies to create a supervised learning machine learning model. The label or outcome variable would be match or not match. In the most naïve sense the dataset would be an outer join or cartesian product of all records in the data sets we're attempting to join. In reality, there is likely a way to reduce that initial number of records, perhaps through blocking on a particular field or only joining if there is some baseline level of similarity. The training set would need to be a subset of these records that are known to be matches through some other means or that have been hand coded as either a match or not.

B. What features would you create for this problem?

A few come to mind:

- Binary variable of exact match or not between matching fields
- Levenstein distance between each of the matching fields. This would be a continuous numerical variable, that might be useful to discretize into bands that indicate no match, close match, exact match, etc.

C. What models would you use?

I would use a regression classifier like SVM or Logistic Regression in this case. Since we want to find records that are exact matches or highly similar on all similar fields, I think that there will be a clear, likely linear separation between matches and non-matches. In reality I would of course train models of many different types, but this would be my starting point.

D. What evaluation metrics would you use?

This is a classification task that would benefit from a confusion matrix. This will give us insight into accuracy, precision, and recall. In this case, it is more important that we capture all true matches than avoid false matches so value recall over precision. This means that we will accept more false positives than false negatives.

E. Would you expect the machine learning solution to work better than exact matching or "fuzzy"/approximate matching rules? Why or why not?

In this case, I am building upon the exact and fuzzy matching rules but not developing the probabilistic model used to generate match results. In this case it's not obvious that machine learning will perform better at record matching because true matches will be such a small proportion of the dataset. This will make training a classifier potentially more difficult and less accurate than a simple rule set.

In thinking further about the probability-based rule set employed in fuzzy matching approaches, learning a tree may be a way to automatically approximate that rule set, which would make both approaches functionally similar.