Questions/Answers

1. If you had the opportunity to build this dataset yourself, what types of things would you have been sure to consider?

Things I would consider if building the dataset for myself:

- Whether their account is being charged to themselves or their company.
- Which courses were they viewing and details about those courses like the authors or the average length of the clips.
- Were any interventions used, like sending them advertisements or special rates?
- Gender/age data

2. Before you begin the analysis, what types of things would you expect to influence the likelihood of whether or not a subscriber renews?

I'd expect more churn in very new and very old customers.
Metrics on activity, like ViewMinPerDay, are probably positively correlated. Many of these variables seem to be probing whether or not the customer is taking a "slow and steady" approach, regularly doing coursework at a maintainable rate.
The significance of CourseCompletionRate probably hinges on the internal definition of course completion, since most users probably jump around. I don't know anyone who's read a textbook from cover to cover.
Variables like CoursePillar, Channel, CountryCode, may be significant in ways I couldn't guess.
TurnAutoRenewOff is a very clear indicator that the customer is leaving.

3. After exploring the dataset, what data cleansing did you do? And, what do you notice that might influence how you approach the analysis and modeling?

Types of data cleaning I did:

- Dates: Converted to a form pandas recognizes as a date, then converted into a numeric variable the classifier can read.
- Errors: 4 observations had negative DaysSinceLastLogin values. This must be an error, so those observations were removed from the dataset.
- Missing values: For the categorical variables, missing values were relabeled as "Null." It's possible that these values were missing for a reason that also affects churn, so we don't want to lose that information. For DaysSinceLastLogin, the only numeric variable with missing values, I didn't want to remove those observations from the dataset and thereby risk introducing bias to the model. So I imputed their value with a very naive guess about what it might be—the median DaysSinceLastLogin.
- High dimensional categorical variables: Categorical variables are unusuable if . I grouped CountryCode into a Continent variable, restricted EmailDomain to the most popular domains and an "other" category, and reduced Company to a binary of whether or not they'd reported a company. The Company variable may be promising for an embedded layer in a neural network, however.

There are over four times as many customers who don't churn than there are those that do, which makes this a moderately imbalanced dataset. I chose a decision tree classifier since those handle imbalanced datasets fairly well without needing resampling techniques. I also tried an AdaBoost model in the hopes that it would perform better on the relatively rare churns, but it was marginally worse.

4. Build a model to predict churn and score subscribers that are coming up for renewal. Tell us how you expect your model to perform in practice.

I would expect the random forest model to have an overall accuracy of 90%, with about an 87% chance of accurately classifying the customers who churn.

5. Describe how your model works, assuming that the audience is a non-technical marketing manager.

This is a random forest model, built from decision trees that look just like you'd expect. At every branch, there's a true/false question about which direction to go, until you end up at which class it predicts you to be in. In the most basic decision trees, each branch is determined by what variable will make the end result the least mixed up so you can have an accurate classifier. The random forest is built by making a lot (~100) of decision trees and then keeping the branches that were most common.

6. What did you learn from your analysis and modeling about the characteristics of the subscribers most at risk of not renewing?

Whether or not the customer currently has their account set to autorenew is the strongest predictor of churn.
There are thresholds for churn around the 6 month and 12 month marks.
Timescale wise, monthly metrics of activity may be better indicators than daily.

7. What recommendations would you make for the use of your model?

This model can be used to target interventions to customers who are most at risk of churning.