Overview Steps

1. Framing the problem. As its in kaggle I'm assuming, the problem definition will be easy to understand.
2. Getting The data
3. Data Exploration to gain insights
4. Data Preprocessing to decipher underlying data patterns and also some ML algorithms wont work accurately without proper preprocessing.
5. Build atleast 5 quick and dirty Models and select best ones
6. Fine tuning the model

Repeat step 3-6 until accuracy get to the desired result.

Problem Framing:

Things to consider:

1. How would I frame this problem?

   - Supervised
   - Unsupervised
   - Online
   - Offline

2. Note down the performance measures to be used? Is it ROC/AUC, RMSE

3. Find out Baseline performance measure. Meaning for classification find out the percentage of class distribution. Or for regression; mean/median.

4. How would I solve the problem manually, assuming human intervention is available.

5. List the assumptions we are making.

Data Collection

As it will be hosted on kaggle I dont think we need to do any data integration/fusion from different databases. Just download and start working.

Data Exploration

Good Practice: Create a copy of the dataset and use a single notebook to do it.

1. Get to know the features

   - Create Data Dictionary Table for future reference.
   - Check the datatypes for each feature
   - % of missing values
   - outliers detection
   - Distribution of each feature (Gaussian, Uniform, log?)

2. If its supervised take a good look at target variable.

3. Visualize the data

    1. Univariate: (Quantitative)
        - Histogram / Density Plots
        - Box Plot
        - Violin Point
    2. Univariate: (Categorical)
        - Frequency Table
        - Bar Plot
    3. Multivariate: (Quantitative vs Quantitative)
        - Correlation Matrix
        - Scatter Plot
    4. Multivariate: (Quantitative vs Categorical)
        - Scatter plot
        - box plot
    5. Multivariate: (Categorical vs Categorical)
        - Contingency Table

4. Figure out the data transformations that might be needed for each features.

Data Preprocessing:

1. Data Cleaning:

    - Remove/Fill in missing values if necessary
    - Fix/Remove outliers

2. Normalize/ Standarization if necessary(if my model uses some kind of distance metrics, we have to use it)

3. Encode Categorical features to numerical and Binning Numerical to categorical if necessary.

4. Feature Extraction

    - Can we decompose a feature to 2 or more features. Like email to username,platform,domain (keep in mind the usefulness)

5. Log/Box-Cox Transformation:

    - Should only be done if our model assumes our features are normally distributes and our features are skewed.

6. Feature Selection:

    - Find out the most important features(use Decision tree's feature importance method). Drop non-contributing features if those features are just noise and not help you to generalize.

Model Building:

1. Building atleast 5 models from different categories(Tree Based, Linear, SVM, Bayesian) with default parameters or change it based on educated guess.
2. Do Cross validation to measure the performance of each model
3. Analyze types of error for each model.
4. Recap the preprocessing steps and change if you think it will improve accuracy.
5. Find out best two/three models.

Fine Tuning:

1. Plot Learning Curve
2. Plot Validation Curve
3. Tune the hyperparameters based on Bias-Variance trade-off and doing visual analysis on the curves you have plotted.
4. Ensemble the best 2/3 tuned models.

Check against Test set!

1. If happy stop,or repeat from Preprocessing changing where necessary!