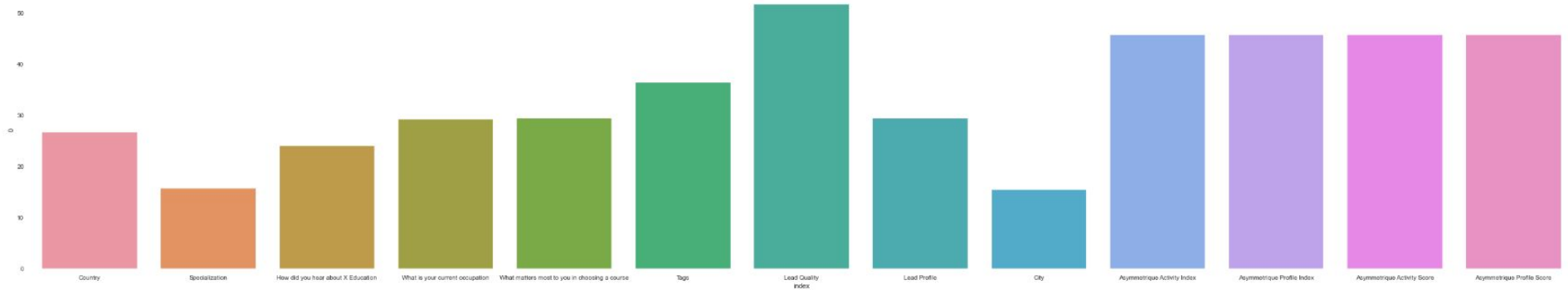
Lead Scoring Case Study

Objective: To assign score to each user that defines the probability of a lead converted to a user

Steps to Solution

- The dataset comprises of 9240 fields and 37 features. Out of that 17 features have null values associated to it.



Preprocessing

- There is absolutely no linkage between feature “Country” and “City” (example: for Country Australia the City was mentioned to be Mumbai) so those were directly dropped.
- Features such as ['How did you hear about X Education', 'What is your current occupation', 'What matters most to you in choosing a course', 'Tags'] are striked off because of abundance of null values and of little significance to the dependent variable i.e. Converted (checked by plotting distribution with dependent variable)

Exploratory Data Analysis

- Used Seaborn library to conduct univariate analysis of features.
- Probability density plot for continuous features (histogram and scatter plot)
- Bar plot for bivariate analysis
- Box plot to check for outliers and distribution across quartiles.
- Sample Distribution: Fig 1.1

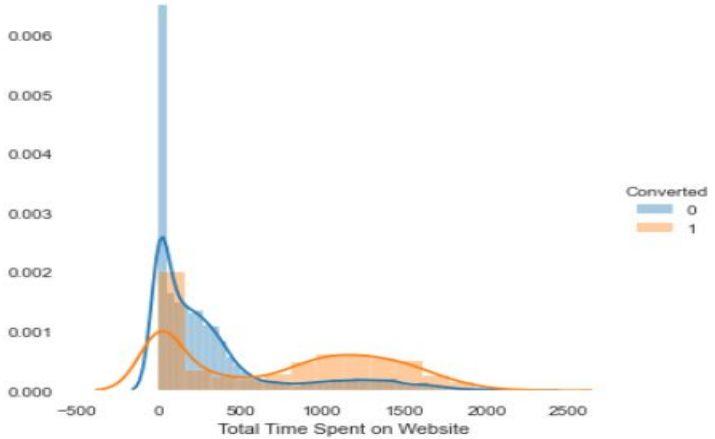


Fig 1.1

Model Building and Accuracy Measures

- Divided Data into three halves:
  - Train
  - Cross Validation
  - Test
- Different Model developed using
  - Logistic Regression
  - Naive Bayes
  - Random Forest (did hyperparameter tuning manually for n_estimators and max_depth)

- Accuracy Measures used:
  - Accuracy score
  - Confusion Matrix
- Compared all the model created using different algorithms
- Found out Random Forest and Logistic Regression to be of greater accuracy, performing almost similar when compared to each other.

Assigning Probability Score

- Logistic Regression was used to predict the probability score for each visitor.
- Sample Snippet: Fig 1.2

```
In [348]: proba=clf.predict_proba(x)

In [350]: final_proba=proba[:,1]*100
          len(final_proba)

Out[350]: 9240

In [351]: df['Prob_Score']=final_proba
```

Fig 1.2

key takeaways

- Understanding the business problem and feature importance to the business is of paramount importance.
- Distribution plot can unveil about the feature a lot.
- Dropping off the features after looking out only the null values can be a disaster, deep dive into the context might find a pattern in missing values too.
- **Cool Feature:** Pandas Profiling Library can automate the process of analysis.