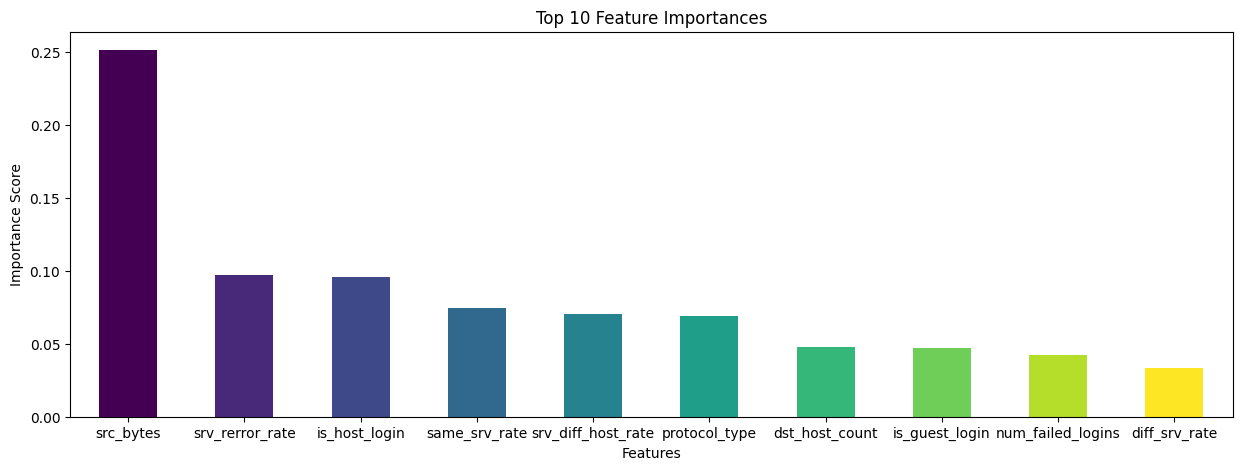
**Feature Importance Analysis**

I selected the Top 10 Feature importance analysis for the Random Forest model and displayed the graph below.



Here’s a breakdown of what these Top 10 features might represent and how they could influence the model's predictions:

### **1. Feature 4: src\_bytes (Importance\_Score: 0.251)**

This feature is by far the most important, with a weight that surpasses the others by a significant margin. **src\_bytes** typically refers to the number of bytes sent from the source to the destination. In network traffic analysis, a high or low number of bytes can indicate different types of behaviour, such as large data transfers (potentially legitimate or malicious) or minimal data exchanges (such as simple pings or scans). Given its high importance, it likely plays a crucial role in determining the nature of the traffic, possibly differentiating between normal and abnormal behaviour.

### **2. Feature 27: srv\_rerror\_rate (Importance\_Score: 0.097)**

This feature represents the percentage of connections that result in errors to the same service. A high error rate might indicate an attempt to connect to a service that is either not available or restricted, possibly signalling unauthorized access attempts or scanning activities. Its significant importance suggests it’s a strong indicator of network misuse or attacks.

### **3. Feature 20: is\_host\_login (Importance\_Score: 0.096)**

This is a binary feature indicating whether a login attempt is to the host system. It plays an essential role in distinguishing between legitimate and illegitimate access attempts. The model places high importance on this feature, likely because unauthorized logins are a primary indicator of security breaches.

### **4. Feature 28: same\_srv\_rate (Importance\_Score: 0.075)**

This feature measures the percentage of connections to the same service. A high rate could suggest repetitive behaviour, such as multiple attempts to access a particular service, which might be indicative of a brute-force attack or automated querying. The model likely uses this feature to detect patterns that deviate from normal usage.

### **5. Feature 30: srv\_diff\_host\_rate (Importance\_Score: 0.070)**

This represents the percentage of connections to different hosts with the same service. An elevated rate might suggest scanning or probing activities where an attacker tries to find vulnerabilities across multiple systems. The importance of this feature suggests that the model uses it to detect such patterns of potential reconnaissance or lateral movement within a network.

### **6. Feature 1: protocol\_type (Importance\_Score: 0.069)**

The type of protocol (e.g., TCP, UDP, ICMP) being used is a fundamental feature. Certain protocols may be more prone to attacks (e.g., UDP for DDoS), and the model likely uses this feature to help classify the type of network activity. Its importance indicates that the choice of protocol is a significant factor in identifying the nature of the traffic.

### **7. Feature 31: dst\_host\_count (Importance\_Score: 0.048)**

This feature counts the number of connections to the same destination host. A high count could indicate a target of interest, either for legitimate reasons or as part of an attack. The model likely considers this feature to help determine whether a host is being unusually targeted.

### **8. Feature 21: is\_guest\_login (Importance\_Score: 0.047)**

Similar to **is\_host\_login**, this binary feature indicates whether the login attempt is from a guest account. Guest logins typically have fewer privileges, and their unexpected use might indicate suspicious behaviour. The model considers this an important feature, likely to detect misuse or privilege escalation attempts.

### **9. Feature 10: num\_failed\_logins (Importance\_Score: 0.042)**

The number of failed login attempts is a classic indicator of brute-force attacks or unauthorized access attempts. The model assigns notable importance to this feature, reflecting its relevance in security contexts.

### **10. Feature 29: diff\_srv\_rate (Importance\_Score: 0.033)**

This measures the percentage of connections to different services. A high rate might indicate varied and potentially malicious probing of different services to identify vulnerabilities. Although it has the lowest importance among the listed features, it still plays a role in identifying complex patterns of behaviour.