**slide:8**

Here's how I would explain the results of the models:

\*\*Model Results:\*\*

Three models were used to predict links in the network, and their performance is summarized below:

1. \*\*Logistic Regression:\*\*

- \*\*AUC-ROC:\*\* 0.9956 – The model is very good at distinguishing between links and non-links.

- \*\*F1 Score:\*\* 0.9563 – A strong balance between precision and recall.

- \*\*Precision:\*\* Between 0.94 and 0.98, depending on the threshold. It means the model correctly identifies links most of the time.

- \*\*Accuracy:\*\* 96% – Overall, the model correctly predicted 96% of the cases.

2. \*\*Random Forest:\*\*

- \*\*AUC-ROC:\*\* 0.9950 – Nearly as good as logistic regression in distinguishing links.

- \*\*F1 Score:\*\* 0.9725 – Very strong balance, slightly better than logistic regression.

- \*\*Precision:\*\* Consistently between 0.97 and 0.98, meaning it’s very good at correctly predicting links.

- \*\*Accuracy:\*\* 97% – This model is also very accurate.

3. \*\*XGBoost:\*\*

- \*\*AUC-ROC:\*\* 0.9957 – The best model at distinguishing links, outperforming the others slightly.

- \*\*F1 Score:\*\* 0.9690 – Shows a good balance, similar to the random forest model.

- \*\*Precision:\*\* Ranges from 0.96 to 0.97, meaning it's also very accurate in predicting links.

- \*\*Accuracy:\*\* 97% – This model is just as accurate as random forest.

**slide:9**

**Feature Importance Analysis:**

In the XGBoost model, we looked at which features were most important in predicting links. The results are:

1. **Adamic-Adar Index:**
   * **Importance:** 0.927862 – This feature is by far the most important. It means the model relies heavily on shared neighbors, especially those with fewer connections, to predict links.
2. **Common Neighbors:**
   * **Importance:** 0.027922 – This feature also helps predict links, but it plays a much smaller role compared to Adamic-Adar.
3. **Jaccard Coefficient:**
   * **Importance:** 0.027589 – This feature also contributes, but like common neighbors, it’s less important than Adamic-Adar.
4. **Preferential Attachment:**
   * **Importance:** 0.016628 – This feature is the least important of all, meaning that the model doesn’t rely as much on the idea that highly connected nodes attract more links.

**slide:10**

**Logistic Regression Coefficients:**

In the Logistic Regression model, the coefficients show how much each feature affects the prediction of links. The results are:

1. **Adamic-Adar Index:**
   * **Coefficient:** 7.149487 – This is the most significant feature. A higher value means that the model strongly relies on shared neighbors with fewer connections to predict links.
2. **Jaccard Coefficient:**
   * **Coefficient:** 4.978354 – This feature also has a positive effect on predicting links, though it’s less significant than Adamic-Adar.
3. **Common Neighbors:**
   * **Coefficient:** -0.265965 – This feature has a **negative correlation**. It means that, according to the model, more common neighbors slightly reduce the likelihood of a link forming.
4. **Preferential Attachment:**
   * **Coefficient:** -0.270951 – This feature also has a **negative correlation**, suggesting that the model doesn’t heavily rely on node degree for link prediction.

**slide:11**

So, **Nodes (2957, 2784)** means we are looking at the potential link between the entity represented by node 2957 and the entity represented by node 2784 in the network. The model predicts the likelihood of a connection (link) forming between these two nodes.  We predicted the chance of links forming between random node pairs using the model. Here are some examples:

1. **Nodes (2957, 2784):**
   * **Prediction:** High chance of a link (0.709228).
   * This means the model thinks there's a good chance these two nodes will connect.
2. **Nodes (3722, 3681):**
   * **Prediction:** Moderate chance of a link (0.576029).
   * These nodes have a moderate chance of connecting.
3. **Nodes (1923, 2505):**
   * **Prediction:** Moderate chance of a link (0.556137).
   * Similar to the previous pair, there's a moderate chance they will connect.
4. **Most Random Pairs:**
   * **Prediction:** Very low chance of a link (<0.05).
   * Most random pairs of nodes have a very small chance of connecting.

**slide:12**

**Node Attributes Insights:**

We looked at key metrics for individual nodes to understand their role in the network:

1. **Node 0:**
   * **Degree Centrality:** 0.0859 – This node has a moderate number of connections.
   * **Clustering Coefficient:** 0.0420 – This node’s neighbors don’t form many connections with each other, so it has low clustering.
2. **Node 2:**
   * **Degree Centrality:** 0.0025 – This node has very few connections.
   * **Clustering Coefficient:** 0.8889 – This node’s neighbors are highly connected with each other, showing high clustering.

**Observation:** Nodes in the network play different roles. Some are more connected, while others have strong relationships with their neighbors but fewer connections.

**slide:13  Challenges and Limitations:**

1. **Technical Challenges:**
   * **Feature Design:** It was difficult to create features that clearly capture the patterns for link prediction.
   * **Model Complexity vs. Interpretability:** Balancing complex models with the need for easy-to-understand results.
   * **Modeling Social Dynamics:** Accurately capturing the complexity of social networks and how connections form.
2. **Data Limitations:**
   * **Anonymized Dataset:** The dataset is anonymized and only from college networks, so it might not apply to other types of networks.
   * **Biases in the Network:** There might be biases in how the original network was built, which could affect predictions.

**slide:14**

**Key Takeaways:**

1. **Project Achievements:**
   * We built accurate models for link prediction with strong performance.
   * Showed how important graph-based features are for understanding social networks.
   * Gained insights into what drives social connections and how they form.
2. **Future Work:**
   * We plan to try more advanced methods like Graph Neural Networks (GNNs).
   * Apply the approach to different datasets to make the model more widely applicable.
   * Improve prediction algorithms for better accuracy and performance on larger networks.

**slide:15**

Conclusion:

This project demonstrates how machine learning models can predict social links within networks, using graph-based features like common neighbors and Jaccard coefficient. It helps us understand how friendships form and evolve, offering valuable insights for building smarter social platforms. The results show that these features are key to capturing network dynamics. Moving forward, exploring methods like Graph Neural Networks (GNNs) can further improve prediction accuracy and scalability, with potential applications in recommendation systems and community building.