Predicting Academic Paper Citation Impact from Peer Review Data

Yixi Zhou, Yifan Huang, Qianyi Gong

April 29, 2025

Agenda

Agenda

Code and Data Availability

Our code and data is available at `https://github.com/XanderZhou2022/ECE204_final`

Project Overview

- **Research Focus:** Exploring the relationship between peer review evaluations and long-term citation impact of academic papers
- **Key Question:** Can early evaluations during the peer review process forecast citation counts?
- **Data Source:** PeerRead dataset with reviews from academic conferences (CONLL 2016, ACL 2017)
- **Approach:** Combine exploratory analysis with predictive modeling to identify patterns in reviewer feedback that correlate with citation outcomes

Why This Problem?

- **Editorial Decision Support:** Help journal editors prioritize papers with high potential impact
- **Understanding Review Process:** Reveal potential biases or blind spots in traditional peer review
- **Author Guidance:** Provide insights on how reviewer feedback relates to eventual research impact
- **Scientific Impact:** Contribute to a deeper understanding of how scientific influence is established and recognized

Agenda

Dataset Overview

- **Source:** PeerRead - A dataset of scientific peer reviews
- **Conferences:** CONLL 2016, ACL 2017
- **Key Components:**
  - Paper metadata (title, authors, etc.)
  - Reviewer scores across multiple dimensions
  - Textual comments from reviewers
  - Acceptance decisions
  - Citation counts (collected via Google Scholar)
- **Sample Size:** Combined dataset of accepted papers from both conferences (around 14.7k)

Review Dimensions in Dataset

- IMPACT - Potential influence on field (evaluated from official reviews)
- SUBSTANCE - Depth of contribution
- APPROPRIATENESS - Suitability for venue
- MEANINGFUL_COMPARISON - Quality of experiments
- SOUNDNESS_CORRECTNESS - Validity
- ORIGINALITY - Novelty of work
- RECOMMENDATION - Overall rating
- CLARITY - Writing quality
- REVIEWER_CONFIDENCE - Reviewer certainty
- PRESENTATION_FORMAT - Delivery format

Figure: Review Dimensions Table (Paper Review Index Breakdown)

Pre-processing Pipeline

1. **Data Collection:**
   - Extracted paper reviews from PeerRead JSON files
   - Used GPT and Google Scholar to collect citation counts
2. **Data Cleaning:**
   - Removed incomplete entries (papers with missing reviews)
   - Converted categorical values (e.g., presentation format) to numerical
   - Handled missing values by dropping incomplete records
   - Checked for outliers in citation counts
3. **Final Dataset:** Clean, structured data ready for analysis

Data Cleaning Steps

- **Converting Categories:** Changed presentation formats to numerical values
  - "Oral Presentation" → 1
  - "Poster" → 2
- **Handling Missing Data:** Removed rows with NA values (could not impute missing review comments)
- **Type Conversion:** Ensured all numerical columns had proper data types
- **Outlier Check:** Used boxplots to identify potential citation count outliers

Figure: Boxplot of Citation Count Distribution (Check the outlier)

Agenda

1. Motivation & Problem
2. Data Acquisition & Cleaning
3. Descriptive Analysis
4. Predictive Modelling
5. Ethical Considerations
6. Conclusion & Future Work

Citation Count Distribution

- Citation counts show a bimodal distribution
- Most papers receive between 50-175 citations
- Some papers achieve much higher citation counts
- The distribution suggests natural groupings of paper impact

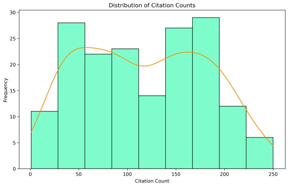

Figure: Distribution of Citation Counts (Section 2.1)

Correlation Analysis

- **Strongest Correlations with Citation Count:**
  - Reviewer confidence has the strongest positive correlation (0.21)
  - Several metrics show weak correlations with citation outcomes
  - Some intuitively important dimensions (like IMPACT) show surprisingly weak correlations with citations

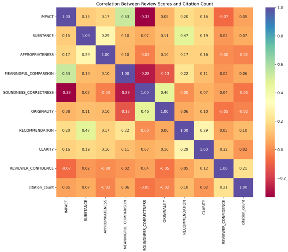

Figure: Correlation Between Review Scores and Citation Count (Section 2.2)

Reviewer Recommendations vs. Citations

- **Key Finding:** Non-linear relationship between reviewer recommendation and citation impact
- **Surprising Pattern:** Papers with lower recommendation scores that still got published often achieved high citation counts
- **Highest recommended papers** don't necessarily receive the most citations

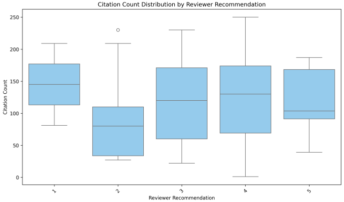

Figure: Citation Count Distribution by Reviewer Recommendation (Section 2.3)

Principal Component Analysis

- Applied PCA to understand dimensionality of reviewer assessments
- First two components explain approximately 43% of variance
- Need at least 5 components to explain 70% of variance
- Suggests reviewer evaluations capture multiple distinct dimensions of paper quality

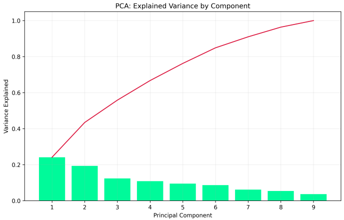

Figure: PCA: Explained Variance by Component (Section 2.4)

Papers in PCA Space

- Visualization reveals no clear clustering of high-citation papers
- Suggests complex relationship between review dimensions and citation impact
- Not easily reducible to a few components

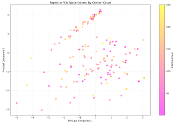

Figure: Papers in PCA Space Colored by Citation Count (Section 2.4)

Clustering Analysis

- Used K-means clustering to identify natural groupings in reviewer assessments
- Optimal number of clusters (k=35) determined using elbow method
- Significant variation in citation impact across different reviewer assessment patterns
- Some clusters show notably higher citation rates

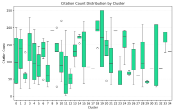

Figure: Citation Count Distribution by Cluster (Section 2.5)

Key Insights from Descriptive Analysis

- **Non-linear relationships:** The relationship between reviewer recommendations and citation impact is not straightforward
- **Multidimensional quality:** Paper quality (as assessed by reviewers) is not easily reducible to one or two factors
- **Reviewer confidence matters:** Confident reviewers may better identify impactful work
- **Limited predictive power:** The relatively weak correlations suggest review scores alone may have limited power in predicting citation impact

Agenda

Prediction Question

Can we predict the future citation count of an academic paper based on reviewer assessments during the peer review process?

- **Target Variable:** Citation count
- **Features:** 10 review dimensions (IMPACT, SUBSTANCE, APPROPRIATENESS, etc.)
- **Machine Learning Task:** Regression problem
- **Evaluation Metrics:** RMSE, R^2, MAE

Model Selection

- **Models Implemented:**
  - Linear Regression - baseline approach
  - Ridge & Lasso Regression - to handle potential multicollinearity
  - Random Forest - to capture non-linear relationships
  - Gradient Boosting - for potentially better predictions
  - Decision Tree - interpretable approach
  - Support Vector Regression (SVR) - for complex patterns
- **Train-Test Split:** 80% training, 20% testing
- **Cross-validation:** 5-fold for hyperparameter tuning

Model Performance Comparison

- **All models showed limited predictive power**
- **Lasso Regression performed best, followed by Support Vector Machine (SVR):**
  - Lowest RMSE and MAE
  - Least negative R^2 score
- **Challenge:** Negative R^2 values across all models indicate they perform worse than simple mean-based prediction

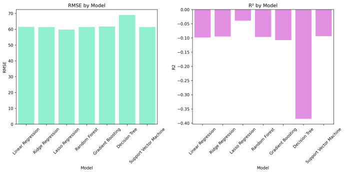

Figure: RMSE and R^2 by Model (Section 3.3)

SVR Model Analysis

- **Best Parameters after Grid Search:**
  - C: Found through grid search
  - Gamma: Found through grid search
  - Kernel: Found through grid search
  - Epsilon: Found through grid search
- **Feature Importance Analysis:**
  - ORIGINALITY and APPROPRIATENESS most influential
  - IMPACT showed surprisingly negative importance
  - RECOMMENDATION had moderate positive importance

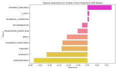

Figure: Feature Importance for Citation Count Prediction (SVR Model) (Section 3.4)

Model Visualization

- **Prediction Challenges:**
  - Model predicts values in narrower range than actual observations
  - Larger errors for papers with higher citation counts
  - Particularly underestimates citation counts for high-impact papers

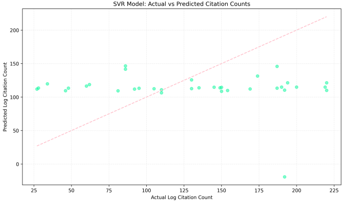

Figure: SVR Model: Actual vs Predicted Citation Counts (Section 3.4)

Part 1 conclusions: Predictive Analysis Insights

- **Limited predictive power:** Peer review assessments alone have limited power to predict citation outcomes accurately
- **Originality matters:** Originality assessments may be more predictive of future impact than other dimensions
- **Non-linear relationships:** The negative R^2 values across all models suggests non-linear relationships between review scores and citation outcomes
- **Prediction challenges:** Models struggle with predicting high citation counts, suggesting that extremely high-impact papers have qualities not fully captured in review scores

XGBoost Model Implementation

- **XGBoost (eXtreme Gradient Boosting)**
  - Advanced implementation of gradient boosting framework
  - Combines multiple decision trees to create a stronger model
  - Well-suited for capturing non-linear relationships
- **Model Configuration:**
  - Objective: Squared error minimization
  - Hyperparameters tuned via RandomizedSearchCV (50 iterations)
  - 5-fold cross-validation for model selection

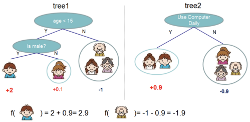

Figure: Tree Ensemble Architecture (Section 3.6)

XGBoost Performance Analysis

- **Model Performance:**
  - Best RMSE: 59.748 (lowest among all tested models)
  - R^2: -0.039 (least negative, but still below zero)
  - MAE: 51.551 (best average error performance)
- **Key Finding:** Despite sophisticated tuning, XGBoost only marginally outperforms simpler models, suggesting fundamental limitations in predictive power of peer review metrics

Model performance comparison (sorted by RMSE):

	Model	RMSE	R2	MAE
2	Lasso Regression	59.747945	-0.039070	51.551050
8	XGBoost (Tuned)	59.850962	-0.042656	52.331596
6	Support Vector Machine	61.305631	-0.093955	52.855627
1	Ridge Regression	61.335642	-0.095027	52.684215
3	Random Forest	61.381788	-0.096675	53.756715
0	Linear Regression	61.427494	-0.098309	52.721620
4	Gradient Boosting	61.686547	-0.107592	53.385936
7	XGBoost (Basic)	68.155326	-0.352068	57.474468
5	Decision Tree	68.971083	-0.384627	52.804762

Figure: Model Performance Comparison (Section 3.6)

XGBoost Feature Importance Analysis

- **Surprising Patterns in Feature Importance:**
  - IMPACT emerged as most influential feature - contrary to correlation analysis
  - REVIEWER_CONFIDENCE and MEANINGFUL_COMPARISON highly important
  - APPROPRIATENESS showed minimal predictive power
- **Contrast with SVR:** XGBoost and SVR models prioritized different features, suggesting complex, model-dependent relationships between review metrics and citations

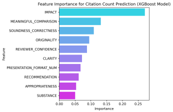

Figure: XGBoost Feature Importance (Section 3.6)

XGBoost Model Visualization

- **Persistent Prediction Challenges:**
  - Predictions clustered between 75-150 citations regardless of actual values
  - Significant underestimation of high-citation papers
  - Large prediction errors across all citation ranges
- **Residual Analysis:** Heteroscedastic pattern with higher variance for higher citation counts

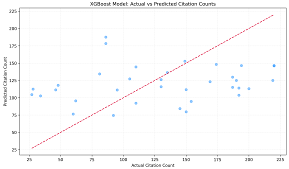

Figure: XGBoost Predicted vs Actual (Section 3.6)

Advanced Non-Linear Methods

- **Explored multiple advanced techniques:**
  - **Neural Networks (MLP)**: RMSE: 69.44, R^2: -0.051
  - **Gaussian Process Regression**: RMSE: 129.48, R^2: -2.65
  - **Polynomial Feature Transformation**: RMSE: 75.03, R^2: -0.227
  - **Stacked Ensemble**: RMSE: 67.35, R^2: 0.0113 (only method achieving positive R^2)
  - **Quantile Regression**: RMSE: 74.04, R^2: -0.195
- **Key Finding:** Even sophisticated non-linear methods struggled, suggesting fundamental limitations in the information content of peer review metrics

Stacked Ensemble Analysis

- **Ensemble Architecture:**
  - Base models: XGBoost, SVR, Random Forest, and Ridge Regression
  - Meta-learners: Ridge, Lasso, ElasticNet, and XGBoost
  - Cross-validated (CV=5) prediction stacking
- **Results:**
  - ElasticNet meta-learner performed best (RMSE: 69.33, R^2: -0.0479)
  - SVR contributed most heavily to ensemble predictions
  - Feature-weighted stacking approach underperformed (R^2: -0.2515)
- **Conclusion:** Stacked ensembles did not significantly outperform individual models, reinforcing that limitations are in the data, not the modeling approach

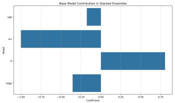

Figure: Base Model Contributions in Stacked Ensemble (Section 3.10)

Part 2 conclusions: Lessons from Advanced Modeling

- **Fundamental data limitations:** Even state-of-the-art modeling techniques failed to achieve strong predictive performance
- **Non-linear relationships confirmed:** Tree-based models consistently outperformed linear approaches
- **Feature importance inconsistency:** Different models emphasized different review dimensions
- **Persistent prediction compression:** All models struggled with the full range of citation outcomes
- **Implications:** Citation impact likely depends on factors beyond peer review metrics:
  - Author reputation and network effects
  - Scientific trends and timing
  - Post-publication promotion and visibility
  - Random elements in how papers gain attention

Agenda

Potential Biases in the Dataset

- **Selection bias:** Dataset includes only accepted papers, creating a truncated view
- **Citation bias:** Citation counts influenced by factors beyond paper quality (author reputation, institutional prestige, etc.)
- **Field-specific norms:** Different academic fields have vastly different citation patterns
- **Temporal effects:** Publications from different years have had different amounts of time to accumulate citations

Implications of Model Use

- **Self-reinforcing biases:** Prediction models could create feedback loops in the publishing system
- **Devaluing innovative research:** Paradigm-shifting research might initially receive mixed reviews
- **Gaming the system:** Authors might optimize for predictive models rather than scientific contribution
- **Disadvantaging certain groups:** Models could perpetuate implicit biases related to gender, institution type, or geographic location

Mitigation Strategies

- **Transparency:** Be clear about model limitations and factors considered
- **Human oversight:** Use models to supplement, not replace, human judgment
- **Regular bias audits:** Continuously monitor and update models for potential biases
- **Field normalization:** Normalize citation predictions by field
- **Diverse metrics:** Consider impact measures beyond citation counts

Agenda

Summary of Findings

- **Weak relationship:** Statistically significant but weak relationships between review scores and citation outcomes
- **Key predictors:** Originality, appropriateness, and reviewer confidence most predictive of citation impact
- **Prediction challenges:** Models struggle to accurately predict citation counts, particularly for high-impact papers
- **Complex relationship:** The relationship between reviewer assessments and citation impact is non-linear and influenced by many factors beyond standard review metrics

Limitations

- **Data constraints:** Dataset limited to accepted papers from specific venues
- **Limited features:** Lack of contextual factors like author reputation and institution prestige
- **Citation count limitations:** Citations are an imperfect proxy for scientific impact
- **Temporal effects:** Analysis doesn't fully account for time to accumulate citations
- **Model limitations:** Standard regression techniques may not fully capture complex relationships

Future Directions

- **Textual analysis:** Analyze the actual text of reviewer comments
- **Multi-modal prediction:** Combine review data with author metrics, institution information, and topic modeling
- **Longitudinal studies:** Track how review assessments predict citation trajectories over time
- **Alternative impact metrics:** Explore relationships with downloads, social media mentions, policy citations
- **Causal analysis:** Use causal inference to isolate influence of specific review dimensions

Data Source & Acknowledgments

- **Data Source:** PeerRead dataset
  - Kang et al. (2018), "A Dataset of Peer Reviews (PeerRead): Collection, Insights and NLP Applications"
  - NAACL 2018
- **Tools Used:**
  - Python with pandas, scikit-learn, matplotlib, seaborn
  - Jupyter Notebooks for analysis

@inproceedings{kang18naacl,
title = {A Dataset of Peer Reviews (PeerRead): Collection, Insights and NLP Applications},
author = {Dongyeop Kang and Waleed Ammar and Bhavana Dalvi and others},
booktitle = {NAACL},
year = {2018}
}

Q&A

Thank you!

Yixi Zhou, Yifan Huang, Qianyi Gong