UNIVERSITY OF TORONTO
DALLA LANA SCHOOL OF PUBLIC HEALTH

A Statlete's approach to predicting TED Talk popularity

Chang Lou, Michela Panarella, Thai-Son Tang, Hongliang Zhang, Yiwen Zhang

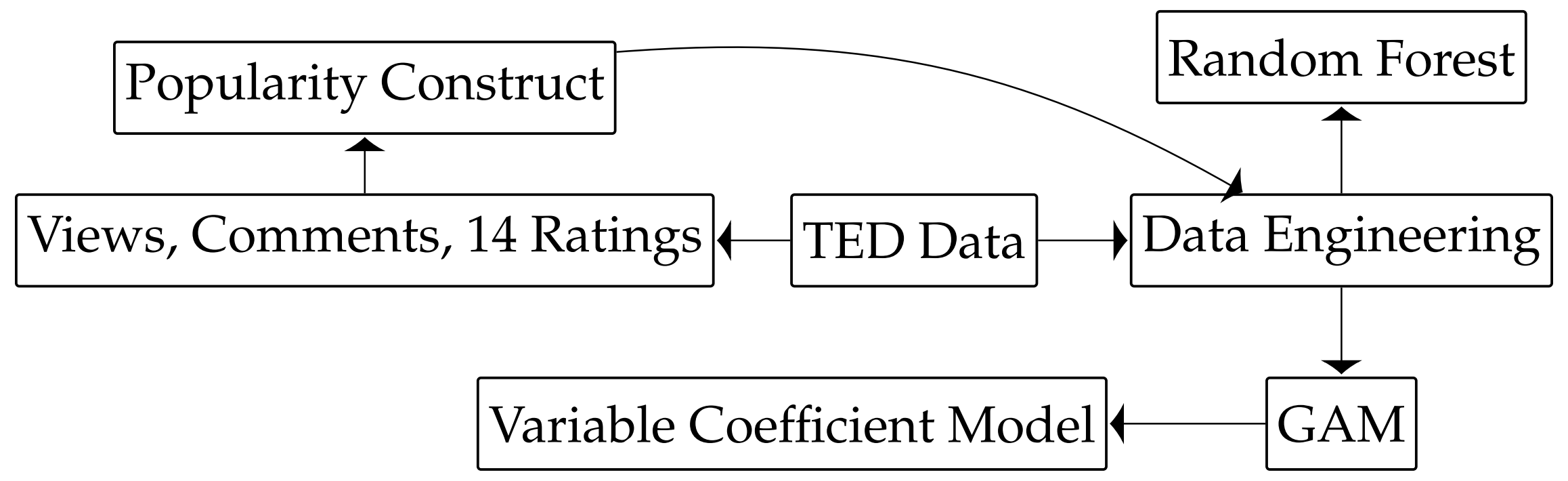
Introduction

Technology, Entertainment, and Design (TED) is an nonprofit organization that shares videos to spread different ideas. Currently, there are thousands of TED talks in over 100 different languages. TED talks are unique and there are limited methods to predict popularity.

Objectives:

1. Define TED Talk popularity.
2. Determine features of TED Talks that predict popularity.
3. Assess if features of TED Talk popularity change with time.
4. Assess if features of TED Talk popularity change with theme.

Methods



Composite Popularity Construct

- **Objective 1:** Factor analysis was performed on the number of views, comments, and 14 ratings, grouping and aggregating the outcome characteristics into **4 factors** (Table 1).
- The sum of the aggregated factors were weighed. The popularity score, Y, of the j^{th} video is the product of the i, j^{th} factor (C_i, j) and weight (ω_i).

$$Y_j = \sum_{i}^{4} \omega_i C_{i,j} \tag{1}$$

Factor	Predictors	Weight
Admirable	Beautiful, Courageous, Inspiring	0.9
Exciting	Views, Fascinating, Ingenious, Jaw-dropping, Funny	0.1
Debatable	Comments, Informative, Persuasive	0.5
Boring	Confusing, Longwinded, OK, Obnoxious, Unconvincing	-0.5

Table 1: The composition and weight of the popularity composite factors.

Data Engineering

Data set was cleaned and predictors used in this study are listed:

Variables			
1	Length of the TED talk video	2	# of subtitled languages
3	# of speakers @ event	4	Season during TED talk
5	**Peak prior-to-talk author popularity**	6	Score based on TED talk tags
7	# of tags linked to video	8	# of related videos to this talk
9	Length of talk title	10	# of words in talk description
11	Words per sentence in talk description	12	**Occupation**: Science/Tech
13	**Occupation**: Art/Lit/Entertain	14	**Occupation**: Sociologist
15	**Occupation**: Activist	16	**Occupation**: Politics/Scholars
17	**Occupation**: Business/Leader	18	Main speaker's # of occupations
19	**Performance video** indicator	20	**Question words** in title indicator
21	TED or independent talk indicator	22	Days since website publication

Table 2: List of variables used in the analysis.

Modeling

Objective 2: Random forest (RF, Eq. 2) and additive models (AM, Eq. 3) were used to capture nonlinear associations between TED Talk characteristics and popularity. RF was optimized by randomly splitting 13 predictors at each node with 4000 trees. AM used thin-spline regression with shrinkage for all continuous covariates in Table 2. Categorical covariates were included as parametric linear predictors. Concurvity was assessed between the AM components to ensure model estimations are stable.

$$Y_j = \frac{1}{B}\sum_{b=1}^{B} f_b(X_{1,j}, ..., X_{p,j}) \tag{2}$$

$$Y_j = \beta_0 + \sum_{i=1}^{p} f_i(X_{i,j}) + \sum_{i=p+1}^{k} \beta_i X_{i,j} + \epsilon_j \tag{3}$$

Objective 3: Varying coefficient models (VCM, Eq. 4) account for nonlinear changes in predictors over time. Cubic regression with shrinkage was applied to all covariates in Table 2. Model concurvity was assessed to ensure the model estimations are stable.

$$Y_j = \beta_0 + \sum_{i=1}^{k} \beta_i(t) X_{i,j} + \sum_{i=p+1}^{k} \beta_i X_{i,j} + \epsilon_j \tag{4}$$

Objective 4: 8 key themes were identified through factor analysis. AMs were utilized on TED Talks in each cluster to identify theme-specific characteristics associated with popularity.

Results

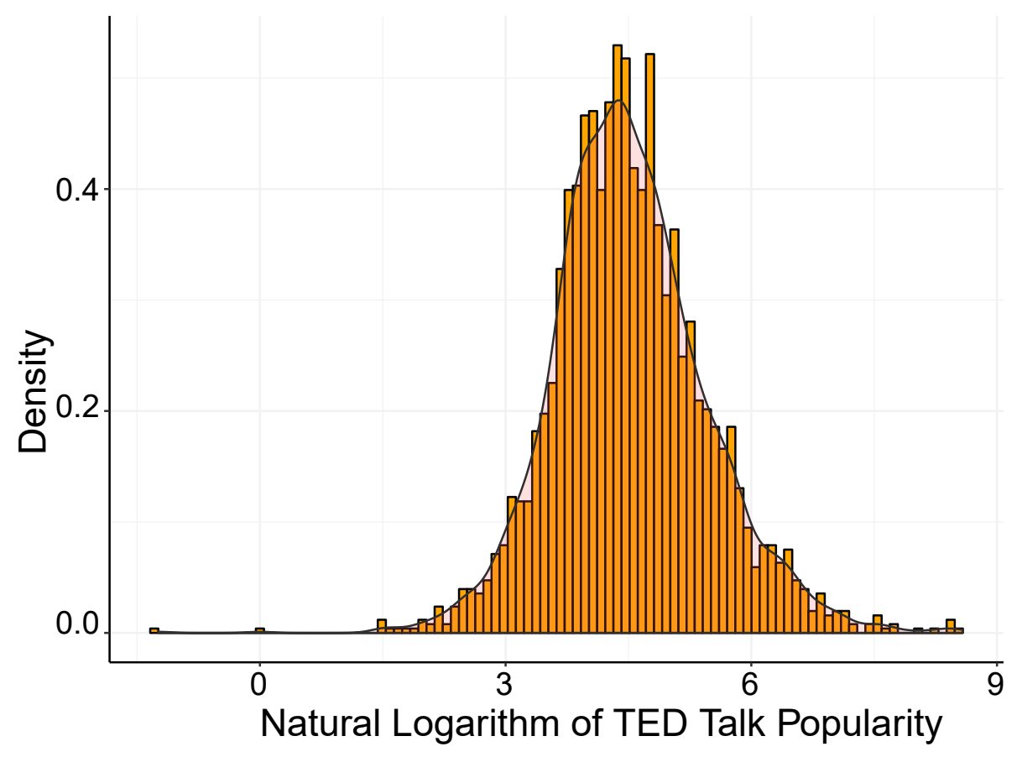


Figure 1: Histogram of the TED Talk Popularity Construct

- The distribution of TED popularity scores was very right-skewed.
- Applying a natural logarithm transformation of the popularity score, we attain a more symmetric distribution for popularity (Figure 1).
- Log-popularity utilized in latter analyses.

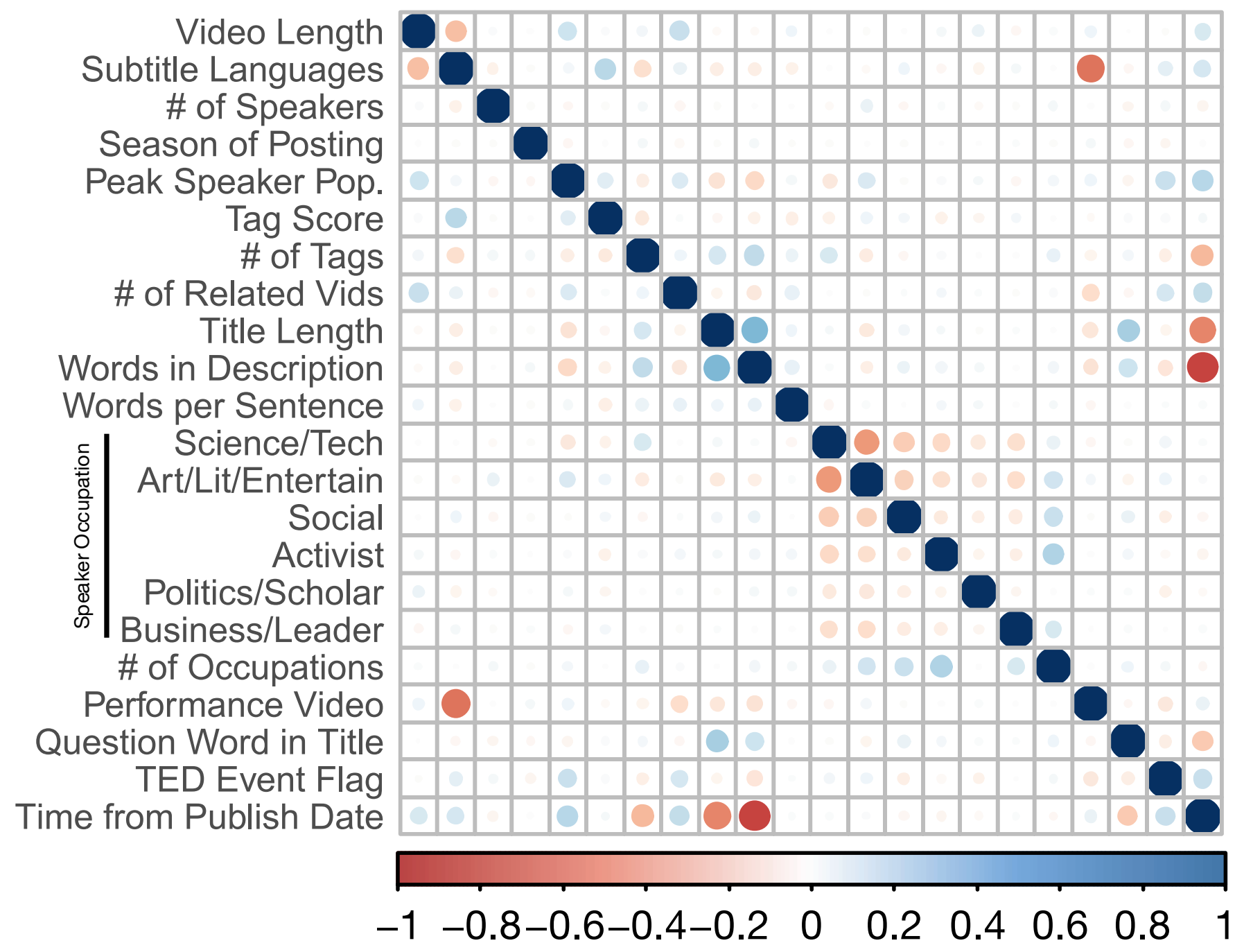


Figure 2: Pearson Correlation matrix of the TED Talk predictors

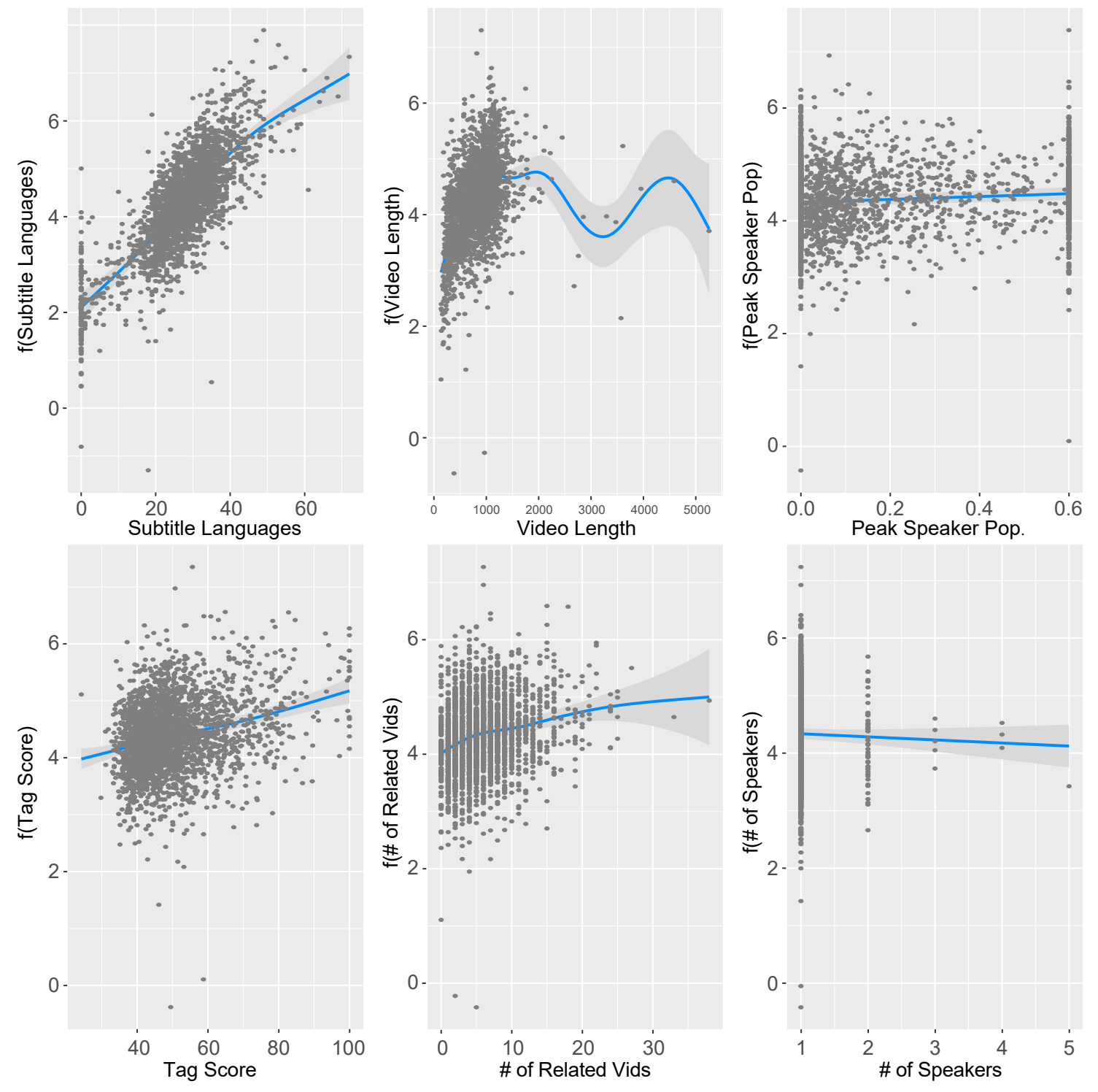


Figure 3: Fitted functions for selected predictors from the generalized additive model.

Figure 3 is the fitted functions for select predictors from the AM. Large values for # of subtitle languages, peak speaker popularity and # of related videos correspond to large values for their corresponding functions. The # of speakers is not significant, with a horizontal line.

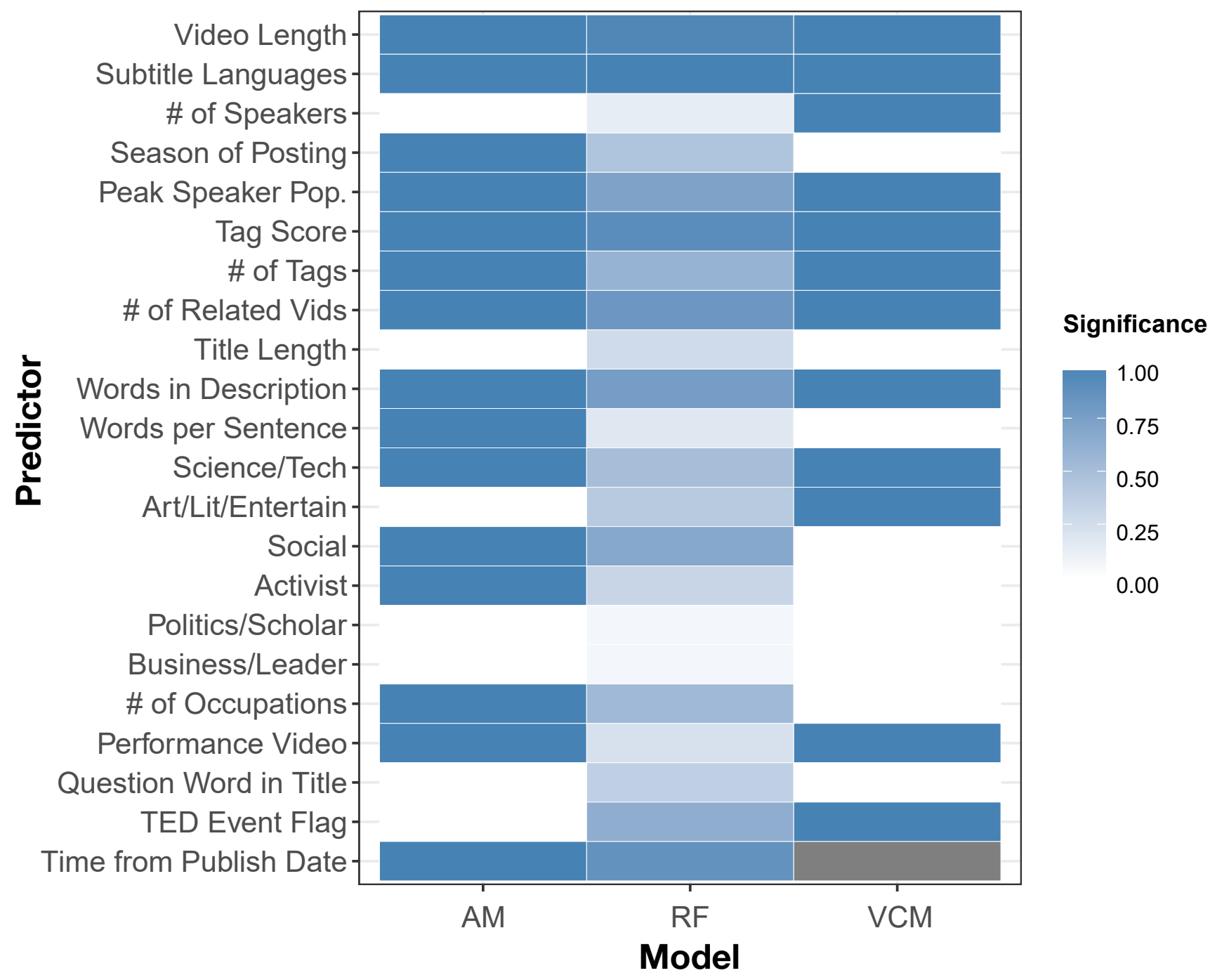


Figure 4: TED talk characteristics associated with popularity based on random forest (RF), additive models (AM) and variable coefficient models (VCM). (Shaded blocks in AM & VCM correspond to $p < 0.05$ and show rankings of variable importance for RF.)

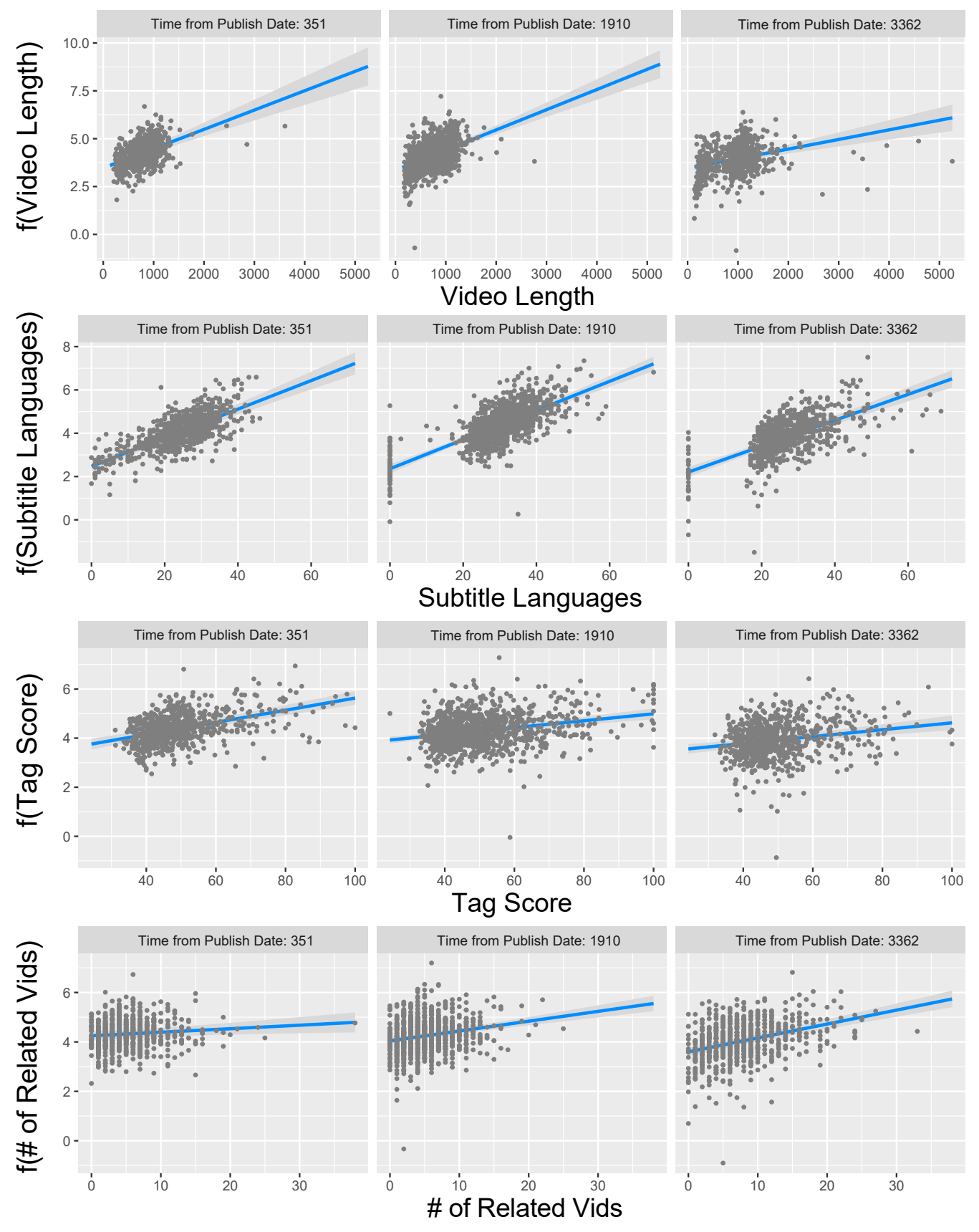


Figure 5: Significant factors (p <0.001) over time.

Figure 5 is how each variable changes at three different time points. Variable importance does change with time when determining TED talk popularity. In Figure 5, longer video length decreases popularity with time, whereas higher subtitle languages and # of related videos correspond to higher popularity.

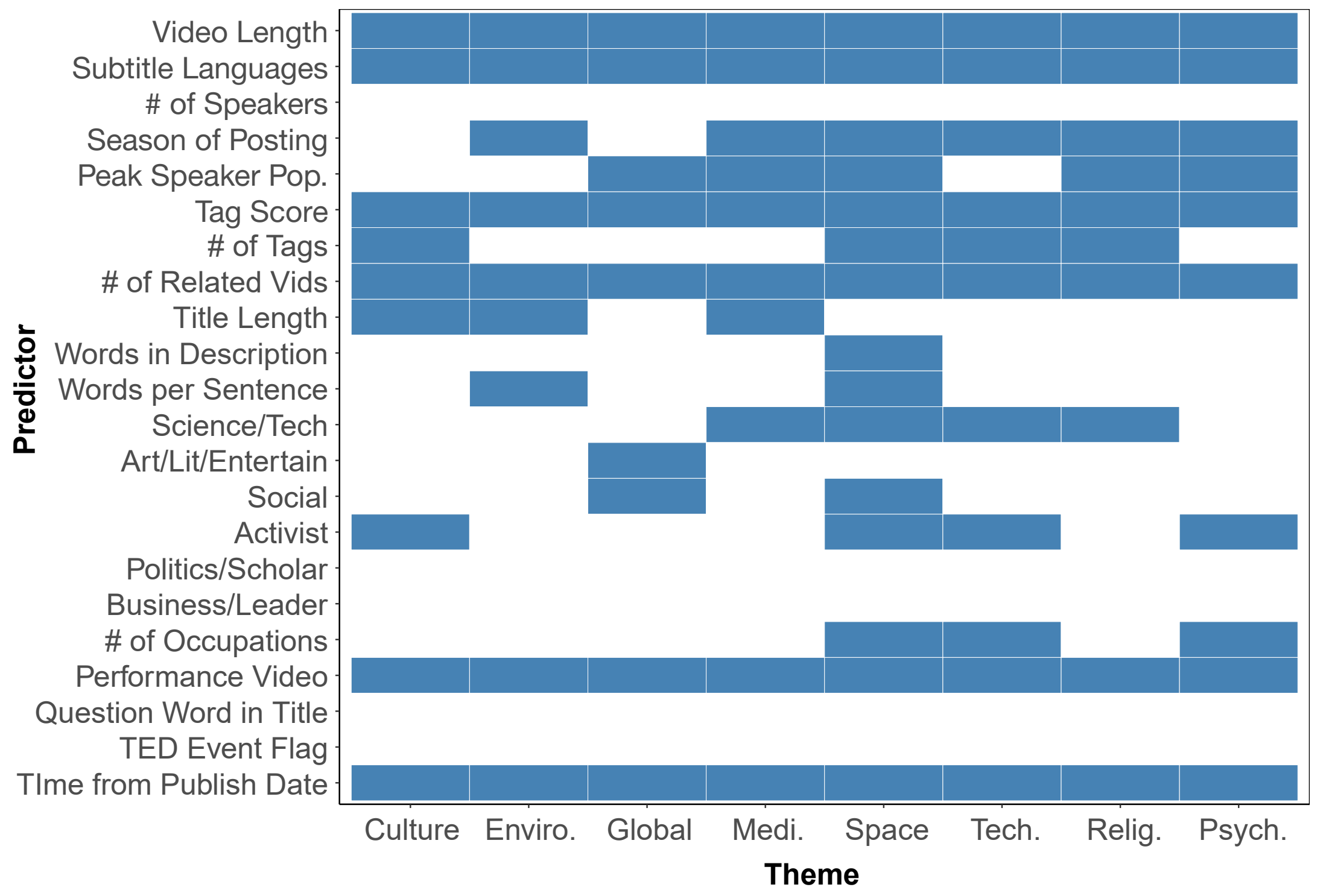


Figure 6: TED talk characteristics associated with popularity in different themes. (Shaded blocks correspond to $p < 0.05$.)

Video length, number of related videos, days since published, and subtitle languages are popularity predictors between all different themes. Description length is associated with popularity in Space while the number of words per description sentence are associated with popularity in Space and Environment. Author popularity has no effect on popularity in Culture, Environment, Medicine, and Technology.

Discussion & Conclusions

- Association doesn't necessarily infer causality. Whether these are variables that could influence future characteristics for TED Talk popularity require implementation of causal inference methods.
- 10-fold cross-validation reveals poor predictability of both GAM ($R^2 = 0.5437, \text{MSE} = 0.4002$) and random forest ($R^2 = 0.5341, \text{MSE} = 0.3969$).
- Dimensionality reduction techniques (PCA, factor analysis) were previously used to create a popularity score, however, forcing 16 outcome variables into 1 dimension does not take into account all the variability. The low R^2 may also reflect this result.
- Our predictions were not great, however we were able to see what does not predict popularity.
- Future work involves extending analysis to multivariate outcomes.

Acknowledgments

We would like to thank our supervisor, Dr. Rafal Kustra, for his invaluable help in the last few months and keeping us on track. As well as, the SSC, the DLSPH Biostatistics Division and Dr. Wendy Lou for providing us the opportunity to participate in this case study.