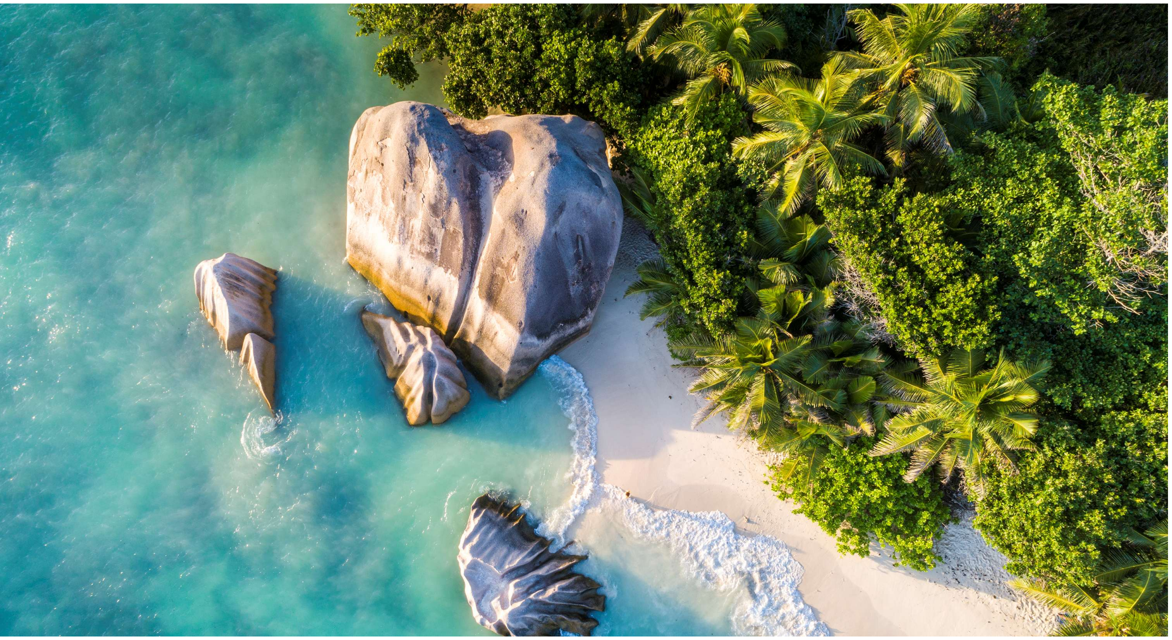
credit

PowerPoint
Stock Images



Project 3:
Web APIs and NLP

Pipistrello
{analytics: LLC}

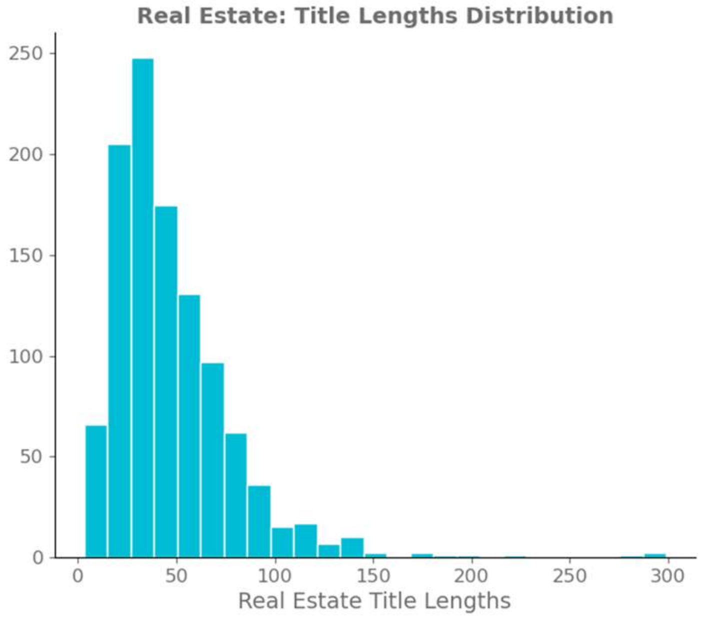
by **Luis Ulloa**
Chief Data Scientist & CEO

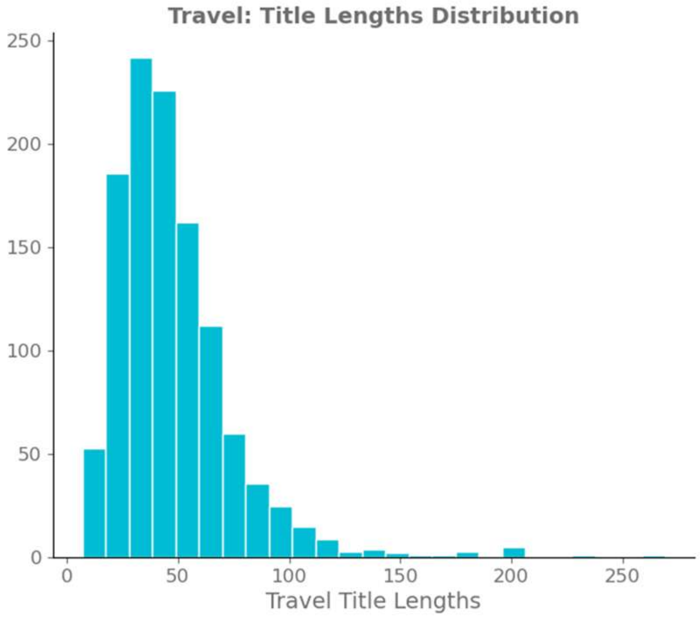
Problem Statement

- **Premise: Can classification algorithms be more accurate than humans?**
  - Scenario: Post titles from two different subreddits: Real Estate and Travel
  - Aim: Train various classification models to correctly guess the subreddit's topic based on the words.

Exploratory Data Analysis

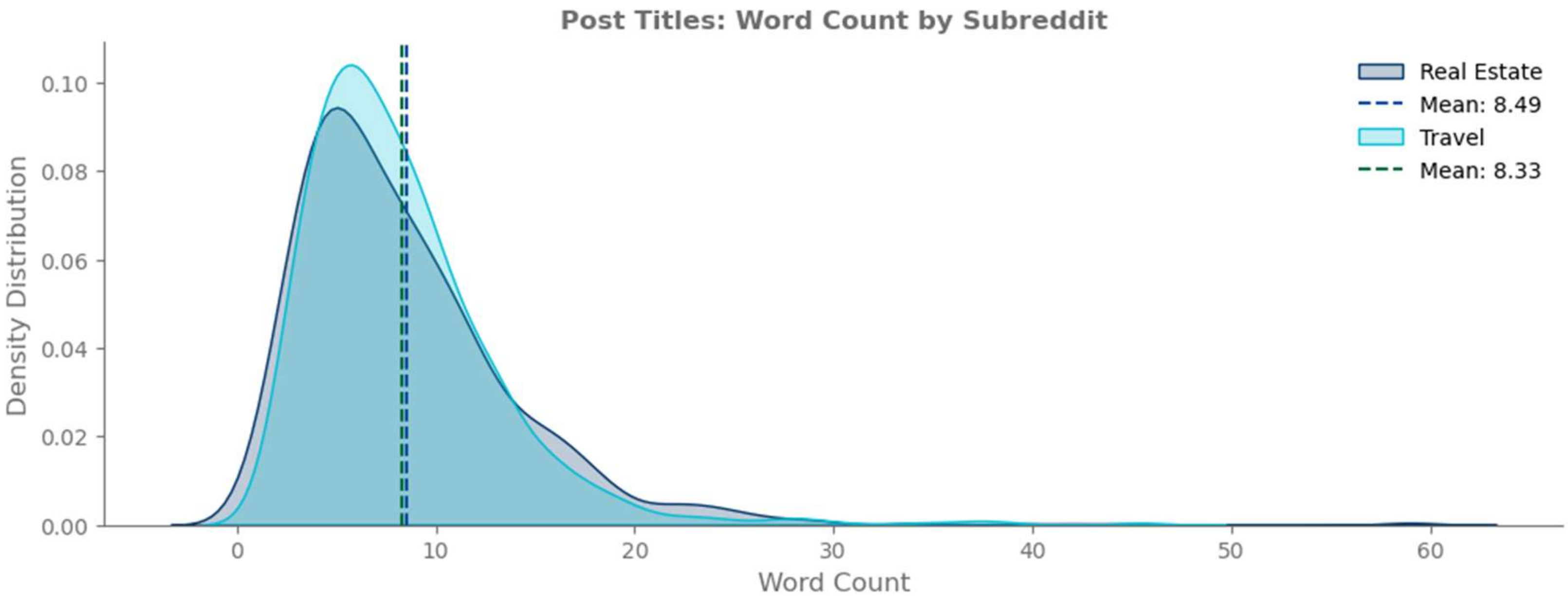
- **Post Titles: Lengths**


Real Estate: Title Lengths Distribution


Travel: Title Lengths Distribution

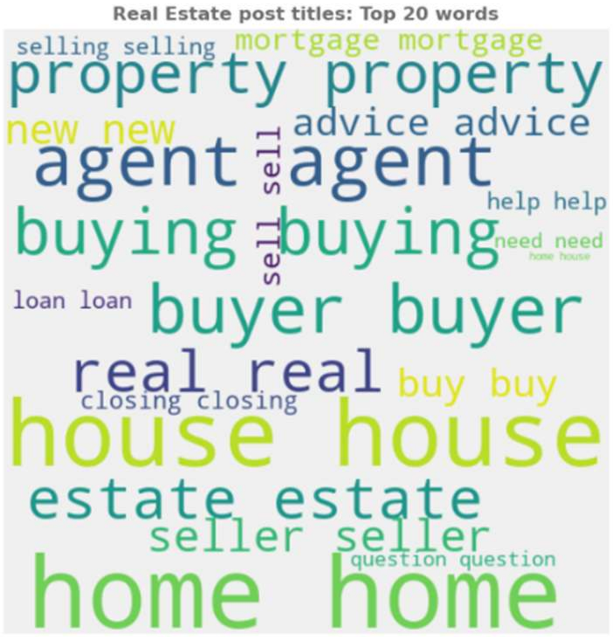
Exploratory Data Analysis

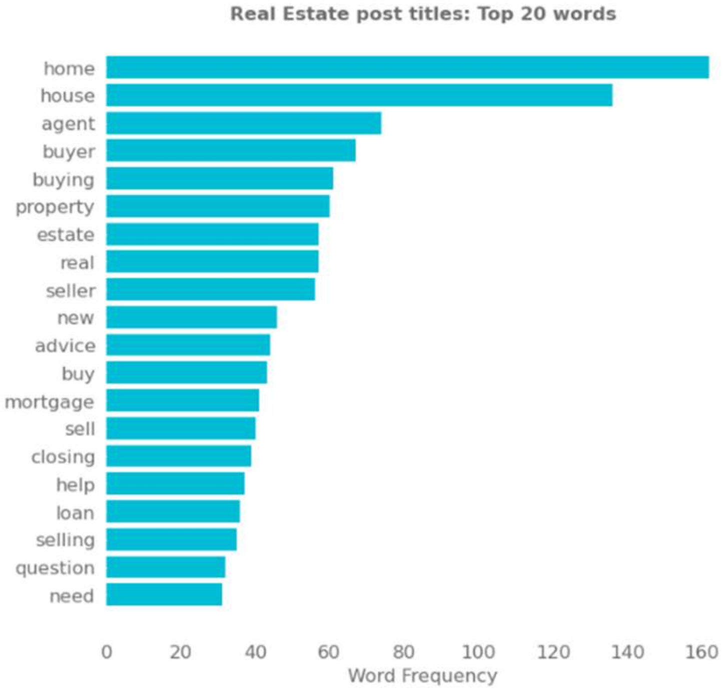
- **Post Titles: Word Counts**



Exploratory Data Analysis

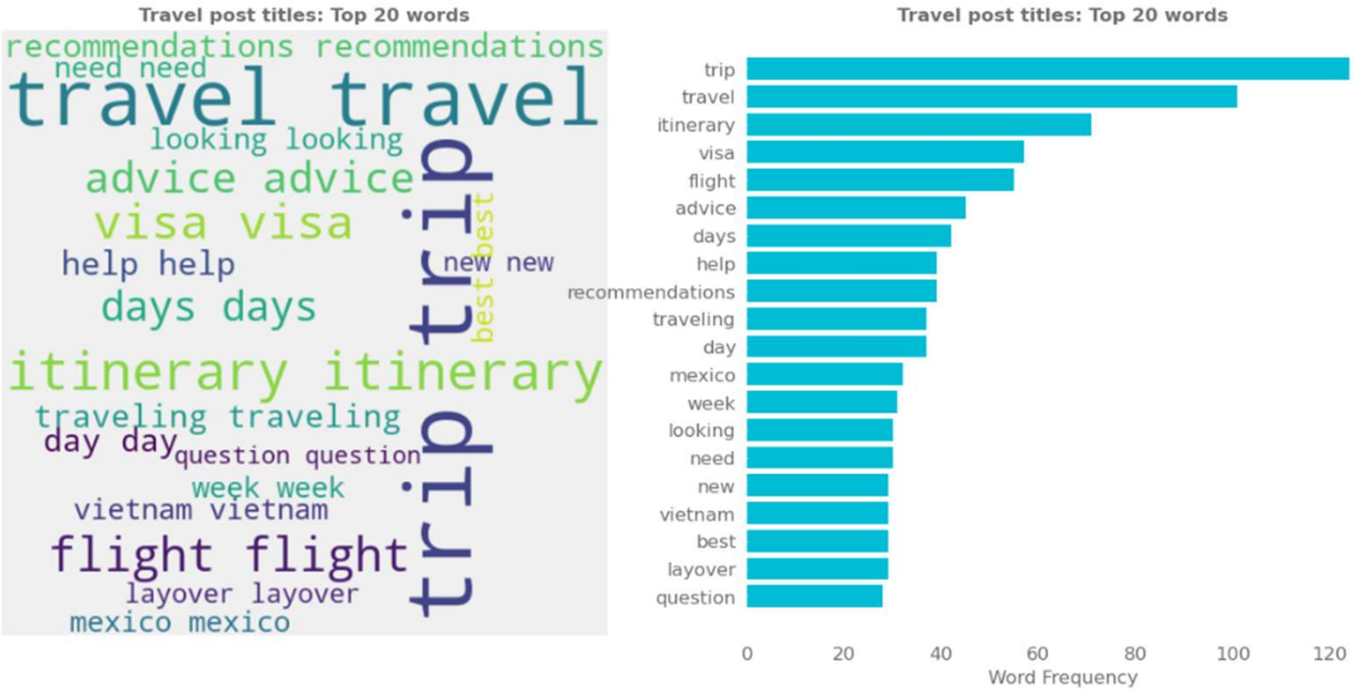
- **Real Estate: Most Frequent Words**





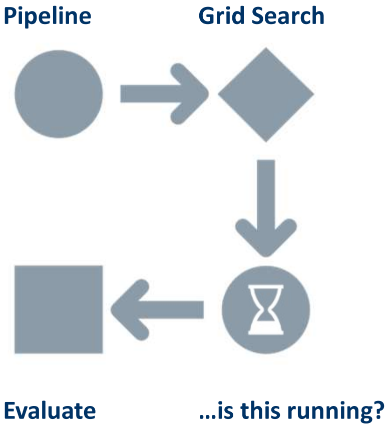
Exploratory Data Analysis

- **Travel: Most Frequent Words**

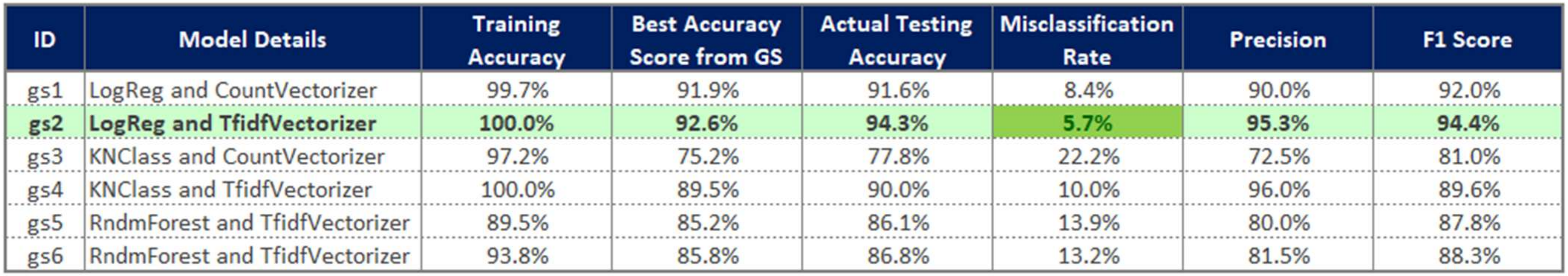


Classification Models

- **Model 1:** Logistic Regression + Count Vectorizer
- **Model 2:** Logistic Regression + Tifid Vectorizer
- **Model 3:** K Neighbors Classifier + Count Vectorizer
- **Model 4:** K Neighbors Classifier + Tifid Vectorizer
- **Model 5:** Random Forest Classifier + Tifid Vectorizer
- **Model 6:** Random Forest Classifier + Tifid Vectorizer

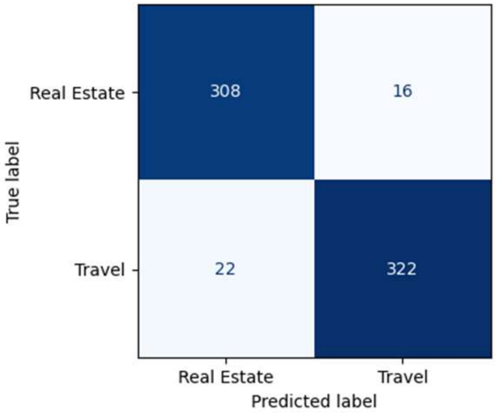


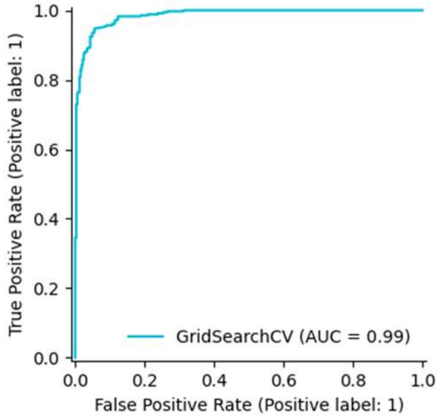
Summary



ID	Model Details	Training Accuracy	Best Accuracy Score from GS	Actual Testing Accuracy	Misclassification Rate	Precision	F1 Score
gs1	LogReg and CountVectorizer	99.7%	91.9%	91.6%	8.4%	90.0%	92.0%
gs2	**LogReg and TfidfVectorizer**	**100.0%**	**92.6%**	**94.3%**	**5.7%**	**95.3%**	**94.4%**
gs3	KNClass and CountVectorizer	97.2%	75.2%	77.8%	22.2%	72.5%	81.0%
gs4	KNClass and TfidfVectorizer	100.0%	89.5%	90.0%	10.0%	96.0%	89.6%
gs5	RndmForest and TfidfVectorizer	89.5%	85.2%	86.1%	13.9%	80.0%	87.8%
gs6	RndmForest and TfidfVectorizer	93.8%	85.8%	86.8%	13.2%	81.5%	88.3%

Baseline Accuracy: 51.5%
Misclassification Rate: 48.5%





Next Steps

- Further fine tuning of hyper parameters to get close to 1% misclassification rate.
- Scrape data from similar subreddits to test model (i.e. Travel vs Travel Hacks)

Questions?