

To help the new grad build machine learning (ML) and data visualization expertise, here are **six project ideas** that combine both skills, along with tools, datasets, and outcomes to highlight on their CV:

## 1. Customer Churn Prediction with Explainability

- **Goal:** Predict which customers are likely to leave a service (e.g., telecom, SaaS).
- **ML Skills:** Classification (Logistic Regression, Random Forest, XGBoost), hyperparameter tuning, feature importance.
- **Visualization:**
  - ROC curves, confusion matrices.
  - SHAP/LIME for model interpretability.
  - Interactive dashboards (Plotly/Dash) to explore churn drivers.
- **Dataset:** [Telco Customer Churn](#).
- **Outcome:** Demonstrate how to turn predictions into actionable business insights.

## 2. Movie Recommendation System

- **Goal:** Build a collaborative or content-based recommender.
- **ML Skills:** Matrix factorization (SVD), cosine similarity, NLP for movie tagline analysis.
- **Visualization:**
  - User-item interaction heatmaps.
  - t-SNE/PCA plots to cluster similar movies.
  - Network graphs for actor/director collaborations.
- **Dataset:** [MovieLens](#).
- **Outcome:** Showcase personalized recommendations and exploratory data analysis (EDA).

## 3. COVID-19 Data Storytelling Dashboard

- **Goal:** Analyze global pandemic trends and vaccination rates.
- **ML Skills:** Time series forecasting (Prophet, ARIMA), clustering to group countries by response.
- **Visualization:**
  - Animated choropleth maps (Plotly/Tableau).
  - Interactive line charts for case trajectories.
  - Heatmaps of vaccination vs. mortality rates.
- **Dataset:** [Johns Hopkins COVID-19 Data](#).
- **Outcome:** Highlight storytelling with dynamic visualizations.

## 4. Credit Card Fraud Detection

- **Goal:** Identify fraudulent transactions using imbalanced data.
- **ML Skills:** Anomaly detection (Isolation Forest, Autoencoders), SMOTE for oversampling.
- **Visualization:**
  - t-SNE for transaction clustering.
  - Precision-recall curves (due to class imbalance).
  - Fraud pattern analysis with seaborn.
- **Dataset:** [Kaggle Credit Card Fraud](#).
- **Outcome:** Emphasize handling imbalanced data and model evaluation.

## 5. Image Classification with CNN Activation Maps

- **Goal:** Classify images (e.g., cats vs. dogs) and explain model decisions.
- **ML Skills:** CNNs (ResNet, VGG16), transfer learning, data augmentation.
- **Visualization:**
  - Grad-CAM activation maps to highlight "important" pixels.
  - Confusion matrices.
  - t-SNE for feature embedding visualization.
- **Dataset:** [CIFAR-10](#) or [Dogs vs. Cats](#).
- **Outcome:** Demonstrate deep learning and model interpretability.

## 6. End-to-End ML Pipeline with Deployment

- **Goal:** Build a full pipeline from data ingestion to deployment.
- **ML Skills:** Model training (scikit-learn), API creation (Flask/FastAPI), Docker.
- **Visualization:**
  - Interactive prediction interfaces (e.g., input sliders in Streamlit).
  - Pipeline architecture diagrams.
- **Dataset:** Any (e.g., housing prices for regression).
- **Outcome:** Show deployment and MLOps familiarity.

## Tips for CV Impact

1. **GitHub Portfolio:** Host code with Jupyter notebooks, clean READMEs, and dependencies.
2. **Blog Posts:** Write Medium/Towards Data Science articles explaining project challenges and solutions.
3. **Tools to Highlight:** Python (pandas, scikit-learn, TensorFlow), visualization libraries (Matplotlib, Seaborn, Plotly), and deployment tools (Heroku, Docker).
4. **Metrics:** Include accuracy, F1-score, AUC-ROC, or business impact (e.g., "Reduced churn by 15% in simulations").

By completing 2-3 of these projects, the grad can confidently showcase their ability to **solve real problems with ML** and **communicate results visually**.